MASTER

An Exploration of a Data-Driven Decision-Making Framework Leveraging the Decision Model and Notation (DMN) Standard: DMN-D3M

Damoiseaux, Jeannot M.B.

Award date:
2023

Eindhoven University of Technology

An Exploration of a Data-Driven Decision-Making Framework Leveraging the Decision Model and Notation (DMN) Standard: DMN-D3M

Author:
drs. Jeannot Damoiseaux

TU/e Supervisors:
dr. ir. Rik Eshuis
dr. Baris Ozkan
dr. ir. Zaharah Bukhsh

Company Supervisors:
ir. Floris Mantz
ing. Koen Lemmens

A thesis submitted in partial fulfillment of the requirements for the
MSc. Operations Management & Logistics
in the Data Intensive Industries track

22 June 2023

Abstract

This study aims to address two key challenges hindering the reliable incorporation of domain knowledge into Machine Learning (ML) projects for data-driven decision-making. The first challenge is the lack of methods for eliciting and validating domain knowledge in ML projects. The second challenge stems from the limited interpretability of black box ML models, which makes it difficult for domain experts to understand and verify the reasoning behind the model's decisions. To overcome these challenges, a novel framework called DMN-D3M is proposed, which integrates domain knowledge and generates interpretable ML models. The framework combines the Decision Model and Notation (DMN) with the Cross Industry Standard Process for Data Mining (CRISP-DM) to provide a unified approach for incorporating domain knowledge throughout the ML project. By formalizing domain knowledge using the Decision Requirements Diagram (DRD), the framework provides a visual and intuitive tool that facilitates discussions and generates insights into the decision-making process, acting as a foundation for subsequent activities. However, in complex scenarios, the interpretability of the ML-based decision tables may become a concern, as they can become challenging to comprehend and validate for domain experts.

Keywords: Data-driven decision-making, Machine Learning, Domain knowledge, Decision Model and Notation (DMN), Decision Requirements Diagram (DRD)

Executive Summary

Problem statement

The ability of ML models to learn from data presents an opportunity for organizations to leverage their historical data to improve their decision-making process (data-driven decision-making). However, the integration of domain knowledge into ML projects is often hindered by two key challenges. First, there is a lack of methods for eliciting and validating domain knowledge in ML projects. Second, black box ML models lack interpretability, making it difficult for domain experts to understand and verify the reasoning behind the system's decisions. Overcoming these challenges is essential to ensure the successful integration of ML as a decision support tool that benefits from the knowledge of domain experts. This leads to the following problem statement:

Problem statement: The lack of methods for eliciting and validating domain knowledge in ML projects, coupled with the limited interpretability of black box models, hinders the reliable incorporation of domain knowledge into ML projects aimed at data-driven decision-making.

Research objective

Based on this problem statement, our research objective is to develop a novel framework for data-driven decision-making that integrates domain knowledge and generates interpretable ML models to enhance the decision-making process. The goal is to create a process framework that integrates a deep understanding of the decisions and their requirements through DMN's requirements level, while producing interpretable ML models at the DMN's decision logic level. This framework is designed for scenarios where the decision logic is not explicitly known in advance and can potentially be derived from historical data using ML algorithms.

Methodology

The research follows the Design Science Research Methodology (DSRM) by Peffers et al. (2007) for Design Science (DS) in Information Systems (IS) research. DS creates and evaluates IT artifacts intended to solve organizational problems (Venable et al., 2016), which aligns with our focus on the development and evaluation of a novel framework for ML in decision-making that incorporates domain knowledge.

The artifact design, development, and evaluation is spread over four phases with each phase building upon the previous phase (Figure 1). In the first phase, we conduct a literature review to explore existing studies that combine ML and domain knowledge through DMN. In the second phase, we define the solution objectives of our artifact and evaluate these solution objectives with practitioners. In the third phase, we develop an initial version of the artifact, based on prior ML and DMN literature, and evaluate it with the same practitioners. In the fourth phase, the framework is demonstrated and evaluated in a rail maintenance setting at Royal HaskoningDHV (RHDHV). In the final three phases, focus groups are conducted in combination with individual questionnaires to evaluate several predefined criteria. For these sessions, we selected 6 practitioners from RHDHV in the role of data analyst (3) or domain expert (3).

Literature review

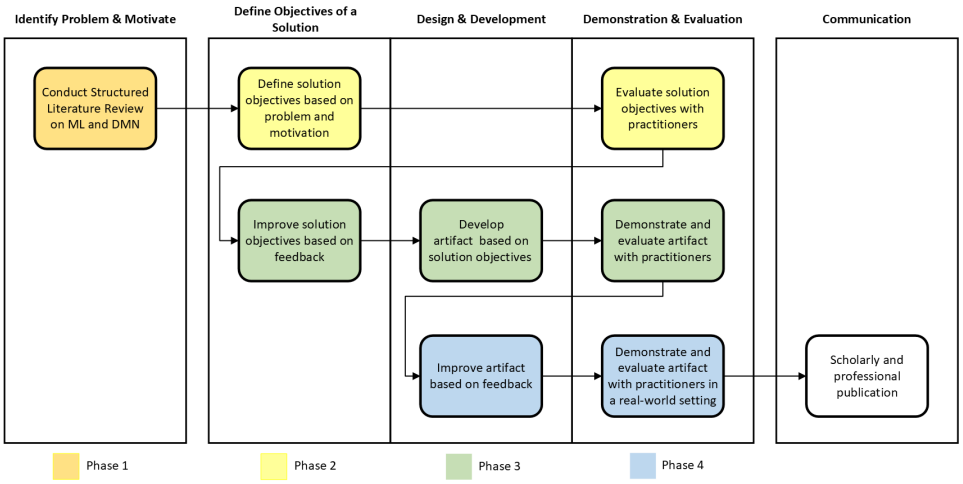


FIGURE 1: Research process model

A systematic literature review focusing on the intersection of DMN and ML reveals a research gap in studies that connect domain knowledge and ML knowledge through DMN, specifically in integrating domain knowledge at the decision requirements level with data-driven approaches to extract the decision-logic level. Existing studies incorporating domain knowledge mainly focus on DMN's decision requirements level, while ML techniques typically focus on decision tables from the decision logic level. By using DMN at both levels, organizations can potentially integrate domain knowledge and improve the interpretability of their ML models.

Framework: DMN-D3M

The proposed DMN-based data-driven decision-making (DMN-D3M) framework consists of five phases (Figure 2). It integrates the two levels of DMN with the Cross Industry Standard Process for Data Mining (CRISP-DM) to create a decision-focused process framework that prioritizes the involvement of domain knowledge. DMN and CRISP-DM are considered complementary methodologies, and the proposed framework systematically integrates both, providing a unified approach to formalize domain knowledge and generate interpretable decision logic. The framework specifically focuses on scenarios where the decision logic is not explicitly known in advance, making ML-based approaches valuable for extracting this logic from the data.

DMN-D3M demonstration

The framework is demonstrated in a rail maintenance setting at RHDHV. This setting involves a condition-based maintenance strategy, which involves yearly inspections to monitor the rail's condition and make preventive maintenance decisions. The rail maintenance context at RHDHV provides a suitable context for this research as RHDHV is aiming to be more data-driven, and less dependent on experts with specialized knowledge.

Three scenarios were explored to extract decision logic in this rail maintenance setting: (1) predicting whether a rail section should be included in the maintenance plan covering the next four years, (2) predicting whether maintenance should be scheduled in the upcoming year, and (3) predicting the exact number of years until the next maintenance. However, scenarios (2) and (3) resulted in unsatisfactory performance measures, leading us to focus on scenario (1) for this

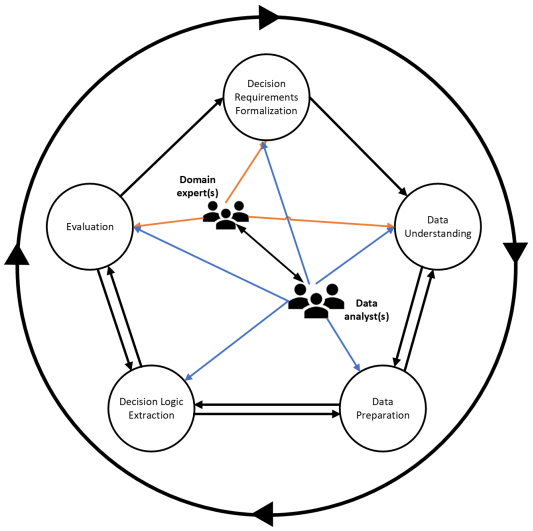


FIGURE 2: DMN-D3M Framework

report. While the extracted decision rules in scenario (1) demonstrate reasonable performance, their practical applicability is currently limited due to several factors. These factors include their inability to take into account cost considerations, changes in the network's overall condition over time and the subjective nature of the underlying condition grades. In the real-world context, decision-makers rely on more detailed observations beyond what is captured in the data, allowing them to make more informed decisions. Therefore, RHDHV is recommended to explore more precise and reliable data collection methods to transition towards a more data-driven process that is less dependent of domain experts.

Discussion

Our findings suggest that the Decision Requirements Diagram (DRD), which specifies DMN's decision requirements level, is valuable for ML projects in a decision-making context. It formalizes domain experts' decision knowledge into a single diagram, which can be presented back to domain experts for validation and serves as a tool for facilitating discussions and generating insights into the decision-making process for data analysts. Additionally, the DRD acts as a foundation for subsequent activities, such as data collection and feature engineering. However, our findings also suggest that the decision table, which specifies DMN's decision logic level, may not always be as understandable and interpretable as previously discussed in research. Particularly, decision tables with a high number of rows and columns increase complexity and reduce comprehensibility. Therefore, the feasibility of a purely DMN-based approach depends on the complexity of the decision logic and the cognitive capabilities of individuals in dealing with such complexity.

Conclusion

In conclusion, this study is the first to integrate domain knowledge through the decision requirements level and employ ML to extract the decision logic level within a process framework for guiding data-driven decision-making. By providing a structured framework, this research contributes to the standardization of incorporating domain knowledge in data-driven decision-making.

Contents

List of Figures

List of Tables

List of Acronyms

AI Artificial Intelligence

BPMN Business Process Model and Notation

CMMN Case Management Model and Notation

CRISP-DM Cross Industry Standard Process for Data Mining

DMN Decision Model and Notation

DMN-D3M DMN-based Data-Driven Decision-Making

DRD Decision Requirements Diagram

DS Design Science

DSMR Design Science Research Methodology

FEDS Framework for Evaluation in Design Science

IS Information Systems

ISDT Information Systems Design Theory

IT Information Technology

ML Machine Learning

NLP Natural Language Processing

OMG Object Management Group

RHDHV Royal HaskoningDHV

SO Solution Objective

WoS Web of Science

XAI eXplainable Artificial Intelligence

1. Introduction

Machine Learning (ML) is a subfield of Artificial Intelligence (AI) that focuses on the development and application of algorithms and models to learn from data and improve their performance on specific tasks without being explicitely programmed. These models have shown outstanding results in various tasks such as image recognition, Natural Language Processing (NLP), recommender systems, and have applications in diverse fields such finance, healthcare, social media and more.

The ability of ML models to learn from data presents an opportunity for organizations to leverage their historical data to improve their decision-making (data-driven decision-making). However, while ML algorithms excel at extracting patterns from data, they may lack the necessary context to fully interpret and comprehend the implications for the decision-making process (Bork et al., 2023). Many studies have focused primarily on the technological aspects of ML, such as data or algorithms, and have neglected the decision aspect (Akter et al., 2019; Chiheb et al., 2019a). This disconnection between the knowledge derived from ML and the actual decision-making process hinders the full potential of data-driven decision-making.

To overcome this limitation, it is crucial to adopt a more holistic and effective approach to data-driven decision-making by integrating a deep understanding of the decisions and their requirements through the incorporation of domain knowledge. In this study, domain knowledge refers to the knowledge of an expert encompassing a comprehensive understanding of a specific decision, including its underlying processes, goals, and requirements. This knowledge often takes the form of business knowledge, as it represents the rules of practice of a specific company (Kopanas et al., 2002).

In our research, we explore the application of the Decision Model and Notation (DMN) as a notation to model and incorporate domain knowledge in ML models. DMN is considered the industry standard for modeling decisions and was introduced the Object Management Group (OMG, 2015). It consists of two levels: the decision requirements level (Figure 3), which captures the high-level dependencies between decision elements, and a decision logic level, which specifies business knowledge through the use of business rules, such as decision tables. By leveraging DMN to incorporate domain knowledge, organizations can potentially enhance the contextual understanding and interpretability of ML models.

1.1 Problem statement

The integration of domain knowledge into ML projects is often hindered by two key challenges. First, there is a lack of methods for eliciting and validating domain knowledge in ML projects. While incorporating domain knowledge into ML through feature engineering is common (Von Rueden et al., 2021; Deng et al., 2020), the process of collecting and incorporating this knowledge is not well-defined and may result in unreliable knowledge. Wagner (2017) noted

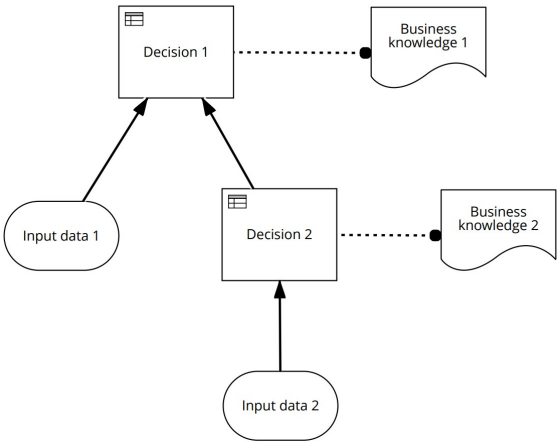


FIGURE 3: Decision Requirements Diagram

that a large number of papers provided no or very little information about the elicitation process, for example, only making cursory references to talking to the domain expert. At the same time, evidence shows that how knowledge is elicited can affect the usefulness of the information (Kerrigan et al., 2021). Similarly, there is a lack of emphasis on validating expert's knowledge in ML projects (Kerrigan et al., 2021). This is especially difficult as domain knowledge is often not quantified and needs to be formalized first (Von Rueden et al., 2021). Validating experts' knowledge can involve presenting it back to them for discussion and confirmation, aggregating it with other experts' responses, or ensuring that it is consistent upon repeated elicitation (Kerrigan et al., 2021). Failing to do so can lead to conflicting or inconsistent knowledge being incorporated into the model, which can negatively affect its accuracy and reliability. Consequently, there is a need for approaches to formalize and validate domain knowledge in ML projects, facilitating more informed decision-making.

The second challenge relates to the limited interpretability of black box ML models. These models operate with internal logic and workings that are hidden from users, making it difficult for domain experts to verify, interpret, and understand the reasoning behind the system's decisions (Montavon et al., 2017). This lack of transparency introduces uncertainty, increases the risk of bias, and decreases trust in the system (Adadi & Berrada, 2018; Carvalho et al., 2019). Consequently, interpretability becomes crucial to ensure that the algorithm performs as expected, particularly in highly regulated domains where decision verifiability is mandatory (Carvalho et al., 2019). Furthermore, providing interpretable ML models enables domain experts to effectively leverage their knowledge and intuition, facilitating a deeper understanding of the model's inner workings. Domain experts can reflect on the model's outputs, analyze its decision-making process, and gain insights into how their domain knowledge aligns with the model's predictions. This reflective process allows experts to identify potential biases, assess the model's limitations, and provide valuable feedback for improvement. Hence, it is essential for ML models to not only be accurate but also to offer interpretability, allowing users to leverage their intuition and reasoning (Varshney et al., 2018).

Overcoming these challenges is essential to ensure the successful integration of ML as a decision support tool that benefits from the knowledge of domain experts. This leads to the following problem statement:

Problem statement: The lack of methods for eliciting and validating domain knowledge in ML projects, coupled with the limited interpretability of black box models, hinders the reliable incorporation of domain knowledge into ML projects aimed at data-driven decision-making.

1.2 Research objective

The primary objective of this research is to address this problem by developing a novel framework for data-driven decision-making that effectively incorporates domain knowledge in ML projects.

Research objective: Develop a novel framework for data-driven decision-making that integrates domain knowledge and generates interpretable ML models to enhance the decision-making process.

The goal is to create a comprehensive process framework that integrates a deep understanding of the decisions and their requirements through the DMN's requirements level, while producing interpretable ML models at the DMN's decision logic level. This framework should provide explicit guidance for executing ML projects within a decision-making context, covering the various stages involved in ML and incorporating domain knowledge as a critical element. Its primary focus is on scenarios where the decision logic is not explicitly known, aiming to uncover the decision logic that is implicitly captured in historical data by leveraging the capabilities of ML algorithms.

This study contributes to the field of Information Systems (IS) research, which is an applied research discipline at the intersection of Information Technology (IT) and organizations (Peffers et al., 2007). The developed framework holds scientific value by advancing the field and providing a novel approach to address the challenges of incorporating domain knowledge and interpretable ML in a decision-making context. Moreover, companies and organizations benefit from adopting such a framework, as it provides a structured approach to execute ML projects in a decision-making context.

1.3 Thesis outline

The proposed framework, called DMN-based data-driven decision-making (DMN-D3M) has been developed using the Design Science Research Methodology (DSRM) by Peffers et al. (2007). This framework integrates DMN and the Cross Industry Standard Process for Data Mining (CRISP-DM), and has been demonstrated in a real-world rail maintenance setting at Royal HaskoningDHV (RHDHV). Its evaluation has been conducted through both ex-ante and ex-post evaluations based on the principles by Sonnenberg & Vom Brocke (2012). Additionally, to ensure a rigorous evaluation strategy, we have employed the Framework for Evaluation in Design Science (FEDS) proposed by Venable et al. (2016) and selected several evaluation criteria for Design Science (DS) in IS from Prat et al. (2014). Furthermore, a comprehensive literature review was conducted as a first step in our DS process, exploring existing studies that combine

ML and domain knowledge through DMN, motivating the subsequent design and development of the DMN-D3M framework.

The report is structured as follows. In Chapter 1: Introduction, the context of this research, the problem statement and research objectives are introduced. Thereafter, the research design is elaborated in Chapter 2: Methodology. Existing research in this direction is then evaluated based on a literature study on DMN and ML in Chapter 3: Phase 1: Literature Review. The three chapters that follow describe the three subsequent phases in our research design: Phase 2: Solution Objectives (Chapter 4), Phase 3: Artifact Design & Development (Chapter 5) and Phase 4: Artifact Demonstration & Evaluation (Chapter 6). Lastly, we discuss the results in Chapter 7: Discussion and draw our final conclusions in Chapter 8: Conclusion and future work.

2. Methodology

2.1 Design Science Research Methodology

In this research, we employ the Design Science Research Methodology (DSRM) by Peffers et al. (2007) for Design Science (DS) in IS research (Figure 4). Design science creates and evaluates IT artifacts intended to solve organizational problems (Venable et al., 2016), which aligns with our focus on the development and evaluation of a novel framework for ML in decision-making that incorporates domain knowledge.

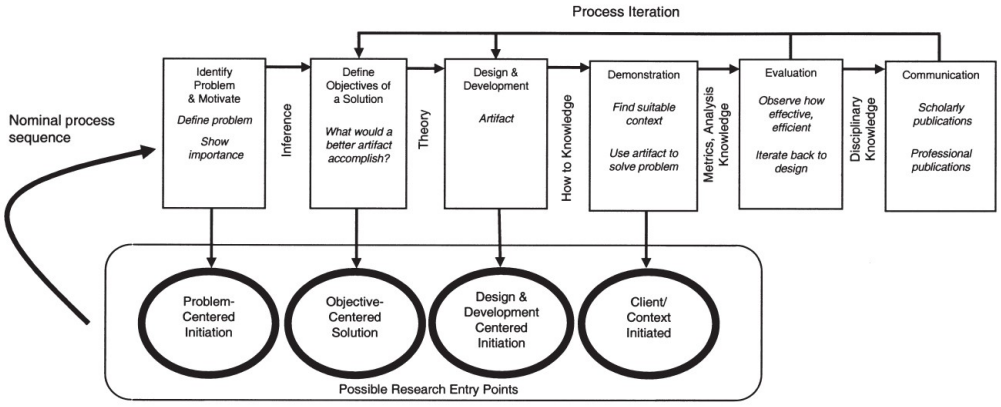


FIGURE 4: DSRM Process Model (Peffers et al., 2007)

This methodology consists of 6 activities, some of which may require multiple iterations to refine the artifact. A problem-centered initiation is the basis of the nominal sequence, which starts with the identification of the main problem and motivation. The second activity involves inferring the objectives of a solution from the problem definition. The problem definition and objectives of a solution are then used to develop an artifact that provides a solution to the problem (activity three). The use of the artifact is demonstrated in the fourth activity by solving one or more instances of the problem. After that, the effectiveness of the artifact as a solution to the problem is evaluated. Lastly, the problem, artifact, design rigor, and its effectiveness are communicated to relevant audiences in the final activity. This process is structured in a nominally sequential order, but researchers may start from of one the first four activities and move outward from there (see other possible research entry points in Figure 4). This research employs a problem-centered approach, building on the problem identification and motivation discussed in the previous chapter.

2.2 The Build-Evaluate Pattern in Design Science

DS in IS comprises of two high-level activities: Build and Evaluate (March & Smith, 1995). These activities can be projected onto the DSMR process model, as illustrated in Figure 5. Evaluations are considered to be crucial in DS and require researchers to rigorously demonstrate the

utility, quality, and efficacy of the artifact (Hevner et al., 2004). Without such evaluations, DS must conclude with only theorizing about the utility of design artifacts, without any empirical evidence (Venable et al., 2016). Typically, such evaluations focus on proving the usefulness of an artifact and less on the artifact design itself (Pries-Heje et al., 2008). However, Sonnenberg & Vom Brocke (2012) argue that evaluations should also consider the importance of design decisions made during the build phase of an artifact. They encourage these decisions to be justified and validated by means of evaluations before an artifact has been put into use, allowing us to make inferences on the usefulness of an artifact, but also its expected suitability, importance, validity and correctness of its design. To accomplish this, Sonnenberg & Vom Brocke (2012) suggested that DS evaluations should be conducted according to the three principles described in the three subsections below.

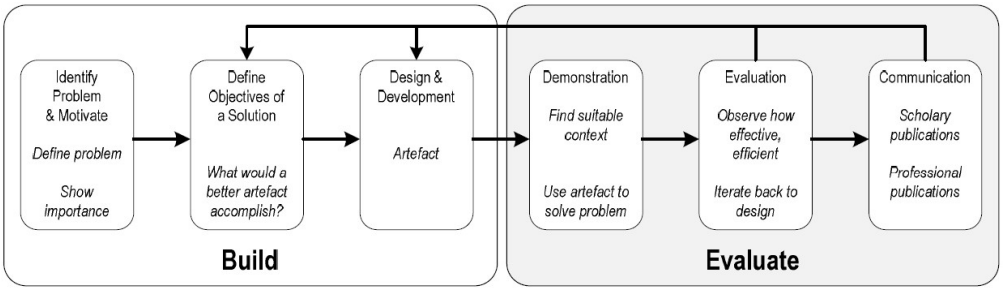


FIGURE 5: The Build-Evaluate pattern projected to DSRM (Sonnenberg & Vom Brocke, 2012)

2.2.1 Interior and exterior modes

First, a distinction should be made between the interior and exterior mode in DS (Figure 6). The interior mode refers to producing prescriptive statements about how the artifact can be designed and developed. On the other hand, the external mode is concerned with producing descriptive knowledge about the artifact, treating the artifact more as a black box and only assessing significant design features for their utility (Gregor, 2009). Rather than relying solely on ex-post evaluations (Table 1) in the exterior mode, Sonnenberg & Vom Brocke (2012) emphasize that it is essential to conduct ex-ante evaluations as part of the interior mode to validate design decisions.

TABLE 1: Evaluation terms [I]

Ex-ante evaluation is "the predictive evaluation which is performed in order to estimate and evaluate the impact of future situations" (Stefanou, 2001, p. 206).	*vs*	*Ex-post evaluation* is an assessment of "the value of the implemented system on the basis of both financial and non-financial measures" (Stefanou, 2001, p. 206).

2.2.2 Documenting design theories

Second, the prescriptive design knowledge in the interior mode should be documented by means of a design theory (Gregor & Jones, 2007). It should document the artifact such that it reveals

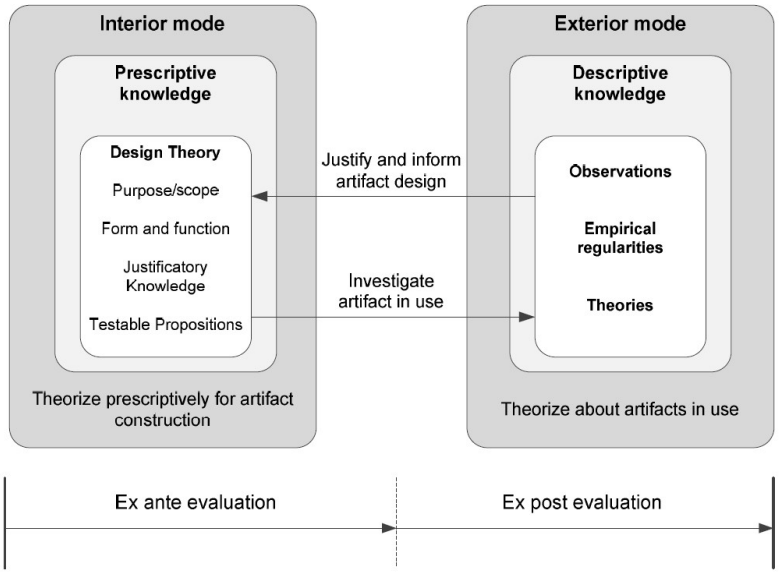


FIGURE 6: Interior and Exterior mode in DS (Sonnenberg & Vom Brocke, 2012)

its purpose, its rationale, its inner structure, the conditions under which the artifact is expected to work, the steps required to actually use the artifact in practice, or testable propositions that can be evaluated in the exterior mode. According to Gregor & Jones (2007) Information Systems Design Theories (ISDT) consists of 8 components (see below). It also connects the interior mode to the exterior mode through components 5, 6, and 8.

1. **Purpose and scope**: Defines the goals and objectives of the artifact, and the context in which it will be used.
2. **Constructs**: Refers to the representations of entities of interest, and the methods and techniques used to construct them.
3. **Principle of form and function**: Includes the principles of form and function that define the structure, organization, and functioning of the artifact.
4. **Artifact mutability**: Specifies the degree to which the designed artifact can be changed or adapted over time.
5. **Testable propositions**: Refers to hypotheses about the artifact to be constructed, which can be validated through testing and experimentation.
6. **Justifactory knowledge**: Comprises the theoretical frameworks, empirical evidence, and practical experience that explain why the artifact might work in a given context.
7. **Principles of implementation**: Includes the methods and techniques used to develop the method.
8. **Expository instantiation**: Refers to the methods and techniques used to assess the usefulness and applicability of the artifact when applied to some reality.

2.2.3 Continuous assessment

Third, the progress of the artifact achieved in the DS process should be continuously monitored. For example, evaluation criteria have to be defined to systematically demonstrate the progress

and to guide evaluation activities (Aier & Fischer, 2011). In the next section, we describe the four-phase approach we adopt in this study to approach such continuity.

2.3 Artifact Design & Development

The artifact design, development, and evaluation is spread over four phases with each phase building upon the previous phase (Figure 1). In the first phase, we conduct a literature review to explore existing studies that combine ML and domain knowledge through DMN. In the second phase, we define the solution objectives of our artifact and evaluate these solution objectives with practitioners. In the third phase, we develop an initial version of the artifact, based on prior ML and DMN literature, and evaluate it with the same practitioners. In the fourth phase, the framework is demonstrated and evaluated in a real-world setting at Royal HaskoningDHV (RHDHV).

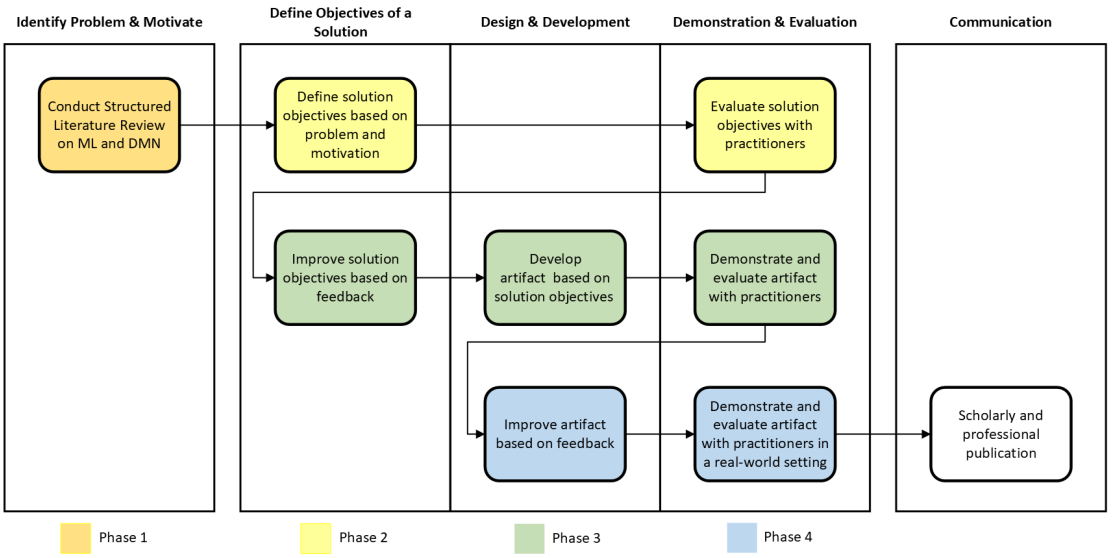


FIGURE 7: Research process model

2.4 Artifact Demonstration

More specifically, the framework is demonstrated in a rail maintenance setting at RHDHV, an international engineering consultancy firm founded and headquartered in the Netherlands. The rail maintenance context at RHDHV provides a suitable context for this research as RHDHV is aiming to be more data-driven, and less dependent on experts with specialized knowledge. Decision-making in rail maintenance heavily relies on such domain knowledge, which aligns with the scope of this study. However, while data is available in rail maintenance, it is unclear how to extract value from it to improve decision-making. RHDHV's motivation for data-driven decision-making in this specific setting is threefold:

1. The maintenance decision-making process currently employed by domain experts is not yet well understood. Although condition data is systematically collected, the conversion

of this data into maintenance decisions is unclear (relying on expert judgement). This is of particular concern as the lack of transferability of such expert judgement combined with difficulty finding experienced replacement staff can threaten the continuity of operations.
2. Asset owners are increasingly seeking predictability, stable costs and safety through well-grounded investment plans and want to move away from solely relying on expert judgement for maintenance decisions.
3. The current condition-based maintenance strategy requires a high inspection frequency, which could potentially be reduced if we can accurately predict maintenance decisions in advance.

To learn from underlying patterns in the data, we have access to 12 years of yearly inspection and maintenance data, which includes both condition grades (e.g. 0-10) and measurements (e.g. 2235 mm). This dataset includes maintenance on almost all rail assets in the network (approximately 45km of rails). Our focus will be on large maintenance, such as lifetime extensions and replacements.

2.5 Artifact Evaluation

In this research, the Framework for Evaluation in Design Science (FEDS) by Venable et al. (2016) is employed to evaluate and develop the artifact. FEDS comprises four steps: (1) explicate the goals of the evaluation, (2) choose the evaluation strategy or strategies, (3) determine the properties to evaluate, and (4) design the individual evaluation episode(s). The implementation for each step in this study is elaborated below. Thereafter, we also discuss the evaluation method and participant selection.

2.5.1 Evaluation goals

We have three main goals when designing the evaluation component of this study. First, we want to ensure scientific rigor. For example, we want to establish that the artifact works in a real-world situation, but also that the artifact causes the observed outcome, and not some confounding independent variable or circumstance. Second, we aim to reduce human social/use risks (risks that the artifact will not fit well into the use or social situation and therefore not work or cause further problems; Venable et al., 2016). Third, we want to ensure a high efficiency of the evaluations by balancing the two aforementioned goals against the available resources for evaluations.

2.5.2 Evaluation strategy and episodes

Our strategy to reach these goals is to conduct an ex-ante formative artificial evaluation (Table 2) in the first and second phase and end with an ex-post summative naturalistic evaluation in the third phase (Table 3). This approach helps us to identify weaknesses and areas for improvement as early as possible, which supports the development of a high quality artifact

and also reduces costs by resolving uncertainties and risks earlier. Furthermore, conducting a formative evaluations as early as possible in the evaluation trajectory helps to reduce human social/use risks (Venable et al., 2016).

Table 2: Evaluation terms [II]

Formative evaluations are used to produce empirically based interpretations that provide a basis for successful action in improving the characteristics or performance of the evaluand (Venable et al., 2016).	*vs*	*Summative evaluations* are used to produce empirically based interpretations that provide a basis for creating shared meanings about the evaluand in the face of different contexts (Venable et al., 2016).
Artificial evaluation includes laboratory experiments, simulations, criteria-based analysis, theoretical arguments, and mathematical proofs (Venable et al., 2016).	*vs*	*Naturalistic evaluation* explores the performance of a solution technology in its real environment, typically within an organization (Venable et al., 2016).

Table 3: Evaluation episodes

Evaluation phase	**When?**	**Why?**	**How?**	**Goal**
Phase 2	Ex-ante	Formative	Artificial	Verify solution objectives
Phase 3	Ex-ante	Formative	Artificial	Verify artifact
Phase 4	Ex-post	Summative	Naturalistic	Explore real-world application

2.5.3 Evaluation properties

In the initial evaluation (Phase 1), we assess the solution objectives based on a relevant subset of the evaluation criteria for DS in IS listed by Prat et al. (2014). The criteria are as follows:

- **Validity**: the degree to which the solution objectives accurately represent the problem and the desired outcomes.
- **Completeness**: the degree to which all necessary aspects are included in the solution objectives.
- **Generality**: the degree to which the solution objectives can be applied beyond a specific domain context.

The artifact evaluations (Phase 2 and 3) focus on a different subset of the criteria listed by Prat et al. (2014):

- **Understandability**: the degree to which the artifact can be comprehended.
- **Completeness**: the degree to which all necessary aspects are included in the artifact.
- **Efficacy**: the degree to which the artifact produces its desired effect (Venable et al., 2012).
- **Generality**: how broad the goal addressed by the artifact is (i.e., the broader the goal addressed by the artifact, the more general the artifact) (Prat et al., 2014).

2.5.4 Evaluation method

In all three phases, focus groups are conducted in combination with individual questionnaires to evaluate these criteria. Focus groups are moderated discussions among selected participants who discuss a topic under the direction of a moderator, whose role is to promote interaction and keep the discussion on the topic of interest (Stewart et al., 2007). Focus groups are considered a relevant and rigorous approach for improving and evaluating design artifacts for several reasons:

1. **Flexibility**: "Focus groups allow for an open format and are flexible enough to handle a wide range of design topics and domains" (Hevner et al., 2010, p.123).
2. **Direct interaction with respondents**: "This allows for the researcher to clarify any questions about the design artifact as well as probing the respondents on certain key design issues" (Hevner et al., 2010, p.124).
3. **Large amounts of rich data**: "The rich data allow deeper understandings, not only on the respondents' reaction and use of the artifact but also on other issues that may be present in a business environment that would impact the design" (Hevner et al., 2010, p.124).
4. **Building on other respondent's comments**: "The group setting allows for the emergence of ideas or opinions that are not usually uncovered in individual interviews. Additionally, causes of disagreement can point to possible problem areas with the proposed artifact" (Hevner et al., 2010, p.124).

However, a limitation of focus groups is that a strongly opinionated member may bias the results and discourage other participants from speaking (Hevner et al., 2010). Therefore, individual questionnaires are conducted before initiating the group discussion to ensure all opinions are discussed. For each criterion, statements are presented, and respondents are asked to indicate the degree to which they agree with these statements on a 5-point Likert scale (Likert, 1932), which then serves as input for the discussion. For example, if a respondent indicates the artifact is not understandable, follow-up questions are asked explore the underlying reasons and potential ways to address it. Lastly, traditional focus group are adapted to the goals of DS in the form of exploratory focus groups (Phase 1 and 2) and confirmatory focus groups (Phase 3) (Hevner et al., 2010).

2.5.5 Participant selection

For these interviews, we selected 6 practitioners from RHDHV in the role of data analyst or domain expert based on the selection criteria in Table 4. More detailed information on these roles will be provided in the artifact description in Phase 3: Artifact Design & Development. Each participant will receive a unique identifier to track their responses, as shown in Table 5.

Data analysts play an essential role in all three phases due to their expertise that allows them to reason about the abstract concepts and terminologies in the artifact, coupled with their high availability. On the other hand, domain experts often lack the background knowledge to grasp the discussed concepts and terminologies, making it challenging for them to provide valuable feedback. This mismatch can lead to domain experts losing interest in the project, which we

want to avoid. Moreover, their availability is often limited due to changing job roles within the company. As a result, data analysts participate in all three phases, while domain experts are involved in the third phase only.

Table 4: Participant selection criteria

Role	**Selection Criteria**
Data Analyst	Education level: Bachelor of Science or higher Experience level: At least 3 years in data analytics or similar (including internship experience)
Domain Expert	Experience level: At least 1.5 years in rail maintenance

Table 5: Participants overview

ID	**Function**	**Role**	**Experience level (years, rounded)**	**Education level**
DA1	Consultant Data & IT	Data analyst	12	MSc.
DA2	Consultant Data & IT	Data analyst	3	BSc.
DA3	Data engineer	Data analyst	7	MSc.
DE1	Asset Manager	Domain expert	37	BSc.
DE2	Project Manager	Domain expert	40	<BSc.
DE3	Project Manager	Domain expert	2	MSc.

3. Phase 1: Literature Review

To address the challenges described in the introduction and to better understand decision-making in organizations, we perform a literature review that focuses on the intersection of DMN and ML. It should be noted that this literature review is a subset of a larger literature study that was performed, focusing on the most relevant parts. Additionally, we added section 3.2.4 in response to a significant publication in the field after the completion of this literature study.

3.1 Methodology

3.1.1 Research questions

Based on the research direction described in the Introduction, we formulate the following main research question:

What are the current methods for using the Decision Model and Notation (DMN) in combination with Machine Learning?

To answer this main question, we divide the question into two parts based on the two DMN levels: decision requirements and decision logic.

1. How does the decision requirements level of DMN interact with ML?
2. How does the decision logic level of DMN interact with ML?

3.1.2 Sources

The research questions described above will be answered through a systematic review of the scientific literature. This review will be based on a search of the reputable scientific databases Scopus and Web of Science (WoS). These databases are considered to be the primary sources for citation data (Mongeon & Paul-Hus, 2016), and are complementary to each other due to their differing coverage (Mongeon & Paul-Hus, 2016; Burnham, 2006). While there may be some limitations to these databases, such as a focus on journals and an over-representation of English-language journals (Mongeon & Paul-Hus, 2016), they are not expected to be major limitations for this review, as the focus will be on journal articles written in English.

- **Scopus** is a database from Elsevier that is considered to be the largest abstract and citation database of research literature and quality web sources (Guz & Rushchitsky, 2009). Journals included in Scopus are reviewed annually to ensure high-quality results.
- **Web of Science**, owned by Clarivate Analytics, is a publisher-independent global citation database with almost 1.9 billion cited references from over 171 million records. It is known for its high level of trustworthiness.

3.1.3 Search terms

To identify relevant literature for this review, a list of search terms was developed based on the main research question (Table 6). These terms include the primary topics "Decision Model and Notation" and "ML," as well as relevant acronyms and variants. For the Decision Model and Notation, the acronym "DMN" was included as a search term. For Machine Learning, the acronym "ML" was used, as well as related variants such as "Artificial Intelligence" ("AI"), "Data Mining," and "Analytics." These terms were chosen to capture a broad range of literature related to the use of DMN in combination with ML.

Table 6: Search terms

Keywords	**Acronyms and variants**
Decision Model and Notation	DMN
ML	ML, Artificial Intelligence, AI, Data Mining, Analytics

It should be noted that the acronym DMN has multiple definitions beyond Decision Model and Notation. To ensure that only relevant literature is included in the review, search results with definitions not related to Decision Model and Notation were excluded. A list of these excluded definitions is provided in Table 7.

To ensure that the review is comprehensive and covers all relevant literature, search terms were also developed for the specific levels of DMN (decision requirements and decision logic). For the decision requirements level (sub-question 1), the search term "Decision Requirements" was used in place of "Decision Model and Notation" and the acronym "DMN" was omitted. This term also captures articles that specifically refer to the Decision Requirements Diagram. For the decision logic level (sub-question 2), the search term "Decision Logic" was used instead of "Decision Requirements", "decision table" table was added as a variant for the decision logic in classification, and both "rule extraction" and "rule mining" were added variants for "ML" in this specific context.

Table 7: Undesired DMN definitions

Acronym	**Definition**
DMN	default mode network, deep material network, dendrite morphological neurons, dynamic memory network, dense semantic matching, deep maxout network, dynamic Markov network, dynamic neural network, disease-associated metabolite network, deep multimodal network, directional mesh network, dynamic manufacturing network, deep matching network

3.1.4 Search queries & results

To identify relevant literature for this review, a search query was developed based on the acronyms and variants listed in Table 6, and the excluded definitions listed in Table 7. This search query was used to search the Scopus and WoS databases for articles in relevant subject areas and in the English language (Table 8). To reduce the number of irrelevant papers, the search was limited to specific subject areas and the language was limited to English as all stakeholders are proficient in this language. The number of results for each database and the total number of unique results is shown in Table 8 as well.

The search queries and results for the specific levels of DMN are shown in Table 9 (decision requirements) and Table 10 (decision logic). Note that, to improve the relevance of the results for the decision logic level, the search query was modified to search only in titles. While this approach may have led to the exclusion of some potentially relevant papers, it was necessary to reduce the high number of irrelevant results (1387 in Scopus and 371 in WoS). It is worth noting that the goal of this review is to provide an overview of the key findings and trends in the existing research, rather than to be exhaustive. Additionally, the use of (reversed) snowballing, a technique that involves following the references cited in papers to find additional relevant papers, can help to ensure that highly relevant papers that may have been missed in the initial search are still identified.

The results from the three search queries are merged into a long list with 67 unique articles.

3.1.5 Selection criteria

To identify the most relevant articles for this review, a selection process was implemented based on the following criteria: language, availability of full text, peer review, and relevance to the research question. We reviewed each article using the following selection and inclusion criteria: 1) the article must be written in English to ensure that all researchers and stakeholders can understand it; 2) the full text of the article must be available online in order to properly evaluate and extract information from it; 3) the article must be peer-reviewed to ensure that it meets good quality standards; 4) most importantly, the article must describe how at least one level of DMN is combined with ML. In addition to these criteria, we will also exclude articles that are not in line with the direction of our research (e.g. those with a focus on incomplete or inconsistent decision tables). We also employed (reversed) snowballing whilst reading the selected papers to identify additional relevant literature. The final short list, as a result of this process, contains 18 unique articles.

3.2 Analysis

In this section, we explore the various methods used to combine DMN and ML. The methods used for the discovery of these DMN models include using domain knowledge, ML and text mining. It is worth noting that there can be overlap between ML and text mining, but for the purpose of this analysis, we categorize an article as utilizing text mining if the data source is text.

Table 8: Main search query

	Scopus	**Web of Science**
Search query	("decision model and notation" OR "DMN") AND ("ML" OR "ML" OR "artificial intelligence" OR "AI" OR "data mining" OR "analytics") AND NOT ("default mode network" OR "deep material network" OR "dendrite morphological neurons" OR "dynamic memory network" OR "dense semantic matching" OR "deep maxout network" OR "dynamic markov network" OR "dynamic neural network" OR "disease-associated metabolite network" OR "deep multimodal network" OR "directional mesh network" OR "dynamic manufacturing network" OR "deep matching network")	("decision model and notation" OR "DMN") AND ("ML" OR "ML" OR "artificial intelligence" OR "AI" OR "data mining" OR "analytics") AND NOT ("default mode network" OR "deep material network" OR "dendrite morphological neurons" OR "dynamic memory network" OR "dense semantic matching" OR "deep maxout network" OR "dynamic markov network" OR "dynamic neural network" OR "disease-associated metabolite network" OR "deep multimodal network" OR "directional mesh network" OR "dynamic manufacturing network" OR "deep matching network")
Search in	Title, Abstract and keywords	Title, Abstract and keywords
Subject area	Computer Science; Engineering; Business Management & Accounting; Decision Sciences	Computer Science, Engineering, Operations Research Management Science
Language filter	English	English
Results	38	15
Unique results	42	

- **Domain knowledge**: This method involves manually gathering knowledge through discussions between the domain expert and the decision modeller until a satisfactory description of the decision is obtained (Etikala & Vanthienen, 2021).
- **Machine Learning**: This method uses historical data, such as case data or event logs, to extract knowledge through ML.
- **Text Mining**: This method involved discovering knowledge through text analysis, which may, or may not, involve ML.

Table 9: Decision requirements search query

	Scopus	**Web of Science**
Search query	"decision requirements" AND ("ML" OR "ML" OR "artificial intelligence" OR "AI" OR "data mining" OR "analytics")	"decision requirements" AND ("ML" OR "ML" OR "artificial intelligence" OR "AI" OR "data mining" OR "analytics")
Search in	Title, Abstract and keywords	Title, Abstract and keywords
Language filter	English	English
Results	9	3
Unique results	11	

Table 10: Decision logic search query

	Scopus	**Web of Science**
Search query	("decision logic" OR "decision table" AND ("ML" OR "ML" OR "artificial intelligence" OR "AI" OR "data mining" OR "analytics" OR "rule extraction" OR "rule mining")	("decision logic" OR "decision table" AND ("ML" OR "ML" OR "artificial intelligence" OR "AI" OR "data mining" OR "analytics" OR "rule extraction" OR "rule mining")
Search in	Title	Title
Language filter	English	English
Results	16	3
Unique results	15	

Our analysis revealed that the majority of the selected articles are in the healthcare and banking industries. This is likely due to the high demand for complex decision-making processes in these fields, as well as the availability of large amounts of data to train ML models. For example, in healthcare, there are many complex decisions that need to be made, such as diagnosis and treatment plans, which can benefit from the use of DMN to model the decision-making process. Additionally, healthcare institutions generate and collect large amounts of patient data, which can be used to train ML models to assist in decision-making. Similarly, in the banking field, there are many decisions that need to be made, such as credit risk assessment, fraud detection, and customer segmentation, and they also have access to a large amount of customer data. A summary of the selected articles, including their discovery method and field is presented in Table 11.

Table 11: Overview of knowledge discovery in selected articles

Article	**Discovery method**		**Field(s)**
	Decision requirements	**Decision logic**	
Bork, Ali & Dinev (2023)	Domain knowledge	Domain knowledge & ML	-
Goossens, De Smedt & Vanthienen (2023)	Text mining	Text Mining	Healthcare
Quishpi, Carmona, & Padró (2021)	Text mining	Text Mining	Healthcare
Arco, Nápoles, Vanhoenshoven, Lara, Casas & Vanhoof (2021)	-	Text Mining	Various
Simić, Tanković & Etinger (2020)	-	ML	Banking, Insurance, Prisoner recidivism
Etikala, Van Veldhoven & Vanthienen (2020)	-	Text mining	Healthcare
De Smedt, Hasić, vanden Broucke & Vanthienen (2019)	ML	ML	Banking
Etinger, Simić, Buljubasic (2019)	-	ML	Iris dataset
Chiheb, Boumahdi & Bouarfa (2019a)	Domain knowledge	-	-
Chiheb, Boumahdi & Bouarfa (2019b)	Domain knowledge	-	-
Car (2018)	Domain knowledge	Domain knowledge	Agriculture
Li, Zhang, Roy & Lee (2017)	Domain knowledge	-	Manufacturing
Servadei, Schmidt & Bär (2016)	Domain knowledge	-	Healthcare
Bazhenova & Weske (2016)	ML	ML	Banking
Bazhenova, Bülow & Weske (2016)	ML	ML	Banking
Lima, Mues & Baesens (2009)	-	ML	Telecom
Mues, Baesens, Setiono & Vanthienen (2005)	-	ML	Banking
Baesens, Setiono, Mues & Vanthienen (2003)	-	ML	Banking
Wets, Vanthienen & Timmermans (1998)	-	ML	-
Our study	Domain knowledge	ML	Rail maintenance

3.2.1 Domain knowledge

In several studies, the utilization of domain knowledge as a discovery method for creating DMN models has been shown to be an effective approach. The majority of these studies focus on the decision requirements level, which allow modellers to make the decision knowledge readily interpretable for all the business stakeholders by highlighting key concepts and relationships.

Both Chiheb et al. (2019a) and Chiheb et al. (2019b) have incorporated the decision requirements level within a theoretical framework to better integrate the decision aspect into ML projects. Servadei et al. (2016) used it to formalize a new decision process that includes a newly trained ML model. Li et al. (2017) used it to form a high-level decision network, visualizing the connections between the data, prediction models, and resulting decisions. In Car (2018), full DMN models (including decision logic) were created based on domain knowledge to investigate its potential use in agriculture decision support. We summarize several advantages of leveraging domain knowledge in the discovery of decision requirements for ML projects:

- **Communication and collaboration**: One of the key benefits of using DMN in ML projects is its ability to facilitate communication and collaboration between domain experts and data professionals. The DRD serves as a bridge between these two groups, allowing them to better understand the business problem or opportunity at hand (Chiheb et al., 2019a; Chiheb et al., 2019b). The relevant knowledge can be visualized and presented to decision-makers in a structured and understandable format, which facilitates the extraction of knowledge and participation among decision-makers.
- **Contextualization**: DRDs and its integration with other industry standards, such as BPMN, allows for contextualization of decisions, showing where ML can be implemented and how they add value (Car, 2018; Servadei et al., 2016). It enables stakeholders to understand the context surrounding a decision, such as the objectives or metrics that are affected by the decision, the required input data, relevant knowledge sources, and other potential decisions that are needed to make the main decision.
- **Documentation**: Documenting the extracted knowledge through DMN can lead to better process execution (Bazhenova et al., 2016) , enable potential reuse at a later stage (Chiheb et al., 2019b; Car, 2018), and serve as guidance or training for new employees.

3.2.2 Machine Learning

In contrast to discovery through domain knowledge, studies employing ML techniques primarily focus on extracting the underlying decision logic in the form of decision tables from historical data. Various algorithms have been utilized in the selected studies, which will be discussed in more detail below.

- **C4.5** (Quinlan, 1993) is a decision tree learning algorithm that converts trained trees into sets of if-then rules and uses the gain ratio to determine the best split at each step of the tree-building process.
- **Trepan** (Craven & Shavlik, 1995) is an algorithm for extracting comprehensible, symbolic representations from trained neural networks by relabeling the training data according to the classifications made by the neural network and then using the relabeled data to train

a decision tree that mimics the behavior of the neural network. These trees have a high level of fidelity to their respective networks while being comprehensible and accurate.

- **Neurorule** (Setiono & Liu, 1996) is a system for extracting symbolic rules from neural networks to understand their behavior. The rule extraction algorithm is able to extract rules that obtain the same accuracy as a pruned neural network, and simpler rules can be obtained by further pruning the network at the cost of accuracy.
- **Neurolinear** (Setiono & Liu, 1997) is a system for extracting compact and comprehensible oblique (as opposed to propositional) decision rules from neural networks, where each condition represents a separating hyperplane given in the form of a linear inequality. e.g.:

$$\textbf{If } 0.84 \text{ Income} + 0.32 \text{ Savings account} \leq 1000 \textbf{ then } \text{Applicant} = \text{bad}$$

- **Layered ensemble model** is an ensemble learning method with complexity reducing techniques proposed by Simić et al. (2020). The overall complexity of the model is configurable, and it suggests a decision with the first decision tree if the purity of the decision leaf is above a preconfigured threshold. If the confidence does not meet the threshold, the decision is passed to the next decision tree until the threshold is met. If the model's confidence never meets the threshold, the final decision is left to the human operator. Etinger et al. (2019) present a method for creating DMN decision tables from a decision tree model.

The decision rules discovered through these algorithms are presented as decision tables, which has several advantages.

- **Interpretability**: Decision tables are widely recognized for their interpretability, which is a crucial factor in their adoption for automating or supporting decision-making (Simić et al., 2020; Lima et al., 2009; Mues et al., 2005; Baesens et al., 2003). They also generate a high level of confidence and are considered easy to use (Huysmans et al., 2011; Martens et al., 2007). The clear and simple representation of decision-making logic makes them easy to understand and use, even by non-technical stakeholders. This is particularly important when validating new knowledge discovered through data before implementing it in the existing business and decision support environment (Mues et al., 2005).
- **Learnability**: The interpretability of decision tables also allows domain experts to identify the key elements in their data that drive decision-making, enabling them to learn and define new strategies (Lima et al., 2009).
- **Maintainability**: Decision table-based systems are also easily maintainable and require minimal effort to modify (Mues et al., 2005). This is a crucial consideration, as ML models can often fail to be successfully integrated into existing business environments due to difficulties with implementation, management, and maintenance of black-box models (Mues et al., 2005).

A minority of studies have aimed to extract complete DMN models, including both decision requirements and decision logic, by using ML techniques. Specifically, if decision trees are used to extract the decision logic level of the DMN model, the decision requirements level can be inferred from the decision logic level. This approach has been attempted using event logs (De Smedt et al., 2019; Bazhenova et al. 2016) and a combination of event logs and process models (Bazhenova & Weske, 2016). Although these studies aim to extract full DMN models, the primary focus remains on discovering the decision logic. Additionally, these studies may

include the decision requirements level as part of the DMN standard, but they do not provide specific argumentation or justification for its inclusion.

3.2.3 Text Mining

More recently, studies have also employed text mining to derive DMN models automatically from text (Arco et al., 2021), primarily in the healthcare domain (Goossens et al., 2023; Quishpi et al., 2021; Etikala et al., 2020). These studies primarily use NLP techniques to extract decision requirements (Etikala et al., 2020), decision logic (Arco et al., 2021) or both (Quishpi et al., 2021). While some researchers argue that manually modeling business decisions is tedious and time-consuming and therefore focus on decision requirements (Etikala et al., 2020), others argue that the creation of DRDs is relatively trivial and may seem redundant, and therefore focus on decision logic (Arco et al., 2021). Some also adopt deep learning techniques to extract full DMN models from text (Goossens et al., 2023). The main advantage of this approach is the ability to create DMN models with less effort and without the need for a deep understanding of the decisions. However, these studies come with several limitations, including the lack of inclusion of other DMN constructs such as knowledge sources and business knowledge models, and the difficulty in handling sentences with multiple outputs (Quishpi et al., 2021) or covering sentences with more than two decision dependency levels (Goossens et al., 2023). Additionally, the identification of relevant sentences from a given business text is often performed manually, which can be problematic in real-world scenarios where sentences with and without relevant decision information are mixed.

3.2.4 Hybrid approaches

After the completion of this literature study, Bork et al. (2023) published a paper that explores the mutual benefits of combining human-driven DMN modeling with the computational power of ML. Their approach, named DMN&ML, uses DMN models to generate ML training data and shows how the trained ML models can enhance human decision modeling by superimposing the feature importances within the original DMN models. By manually modeling DMN and generating datasets that conform the valid input sets and decision rules, they analyze the significance of input features on the decision outputs. This analysis helps in identifying redundant and unnecessary input parameters, which supports the simplification and refactoring of the decision logic. Additionally, this modelling process may provide insights into the explicit and implicit domain knowledge on which the decision logic is based.

3.3 Discussion and conclusion

We found that the combination of DMN and ML has the potential to be a powerful approach for data-driven decision-making. However, there is currently a lack of studies that integrate both domain knowledge and ML approaches through the two levels of DMN in a unified manner. Studies that incorporate domain knowledge mainly focus on DMN's decision requirements

level, which has several advantages such as improved communication and collaboration, better contextualization of the decision-making and documentation of the extracted knowledge. On the other hand, studies that employ ML techniques typically focus on decision tables from the decision logic level, which also has several advantages such as interpretability, learnability and maintainability. More recently, a few studies have also started using text mining to automatically discover DMN models from texts, thereby creating DMN models with less effort and without the need for a deep understanding of the decisions. However, limitations such as the need for manual selection of relevant sentences and their limited capacity to handle the complexity within these sentences restricts their applicability in most real-world scenarios.

Only very recently, Bork et al. (2023) made the first contributions towards combining manual DMN modeling with the computational power of data-driven approaches (ML). They confirmed such hybrid approach is lacking entirely in research, and aimed to enhance human decision modeling by superimposing the feature importances within the original DMN models. However, it is important to note that such approach assumes assumes prior knowledge of the decision logic and therefore still heavily relies on human modeling.

In conclusion, a research gap has been identified in studies that connect domain knowledge ánd ML knowledge through DMN, specifically in the integration of domain knowledge at the decision requirement level with data-driven approaches for extracting the decision-logic level. By using DMN at both levels, organizations can potentially integrate domain knowledge and improve the interpretability of their ML models. Therefore, we believe that further research in this area has the potential to yield significant benefits for organizations seeking to leverage their data to make better decisions.

4. Phase 2: Solution Objectives

As part of the initialization and development of our artifact, we first determine its solution objectives (SO). These objectives are inferred from the problem statement and objective of this research as discussed previously and will guide the design and development of our artifact.

4.1 Define Objectives of a Solution

The first objective of the artifact aligns with three out of the four evaluation criteria defined in 2.5.3: Evaluation properties: understandability, completeness and generality (Prat et al., 2014). Among these properties, understandability plays a crucial role as it enables users to quickly grasp the underlying concepts, principles, and components of the artifact, leading to a more efficient and effective application. Additionally, completeness is important as it ensures that the artifact encompasses all the necessary features and functionality for an effective application. Moreover, generality is a critical property that ensures the usability of the artifact across different applications. These three properties are also interconnected, as an incomplete artifact may hinder understandability, and a generic artifact simplifies complex implementation details, thereby enhancing understandability.

- **SO1**: The artifact provides an understandable, complete and generic process framework for data-driven decision-making.

The subsequent solution objectives align with the fourth evaluation criterion: efficacy (Prat et al., 2014). The second solution objective of the artifact addresses the structured incorporation of domain knowledge. It emphasizes the importance of formalizing and aggregating domain knowledge from various experts, such that it can be presented back to the experts for further discussion and validation. This iterative process ensures the accuracy and relevance of the incorporated knowledge, directly addressing the specific aspect of domain knowledge incorporation stated in the problem statement.

- **SO2**: The artifact describes how to formalize, aggregate and validate domain knowledge relevant for decision-making.

The third solution objective of the artifact focuses on enhancing the interpretability of the resulting ML model. By prioritizing interpretability, we aim to create a model that can be easily understood by its users, which enables them to trust and validate the decisions made by the model (e.g., Adadi & Berrada, 2018; Carvalho et al., 2019). Building upon the advantages identified in our literature review, our aim is to develop decision logic that is interpretable, but also learnable and maintainable. The interpretability allows domain experts to identify the key elements in their data that drive decision-making, enabling them to learn and define new strategies (Lima et al., 2009). By focusing on maintainability, we aim to develop a model that can be updated over time as new data and insights become available. This is crucial as black-box models often pose challenges in terms of implementation, management, and maintenance, hindering their successful integration into existing business environments (Mues et al., 2005).

These objectives contribute to the overall transparency and effectiveness of the ML model in supporting decision-making.

- **SO3**: The artifact generates decision logic that is interpretable, maintainable, and learnable.

The fourth solution objective of the artifact emphasizes the integration of domain knowledge into ML projects and the utilization of interpretable ML models to establish a shared understanding between domain experts and data analysts. By incorprating domain knowledge, we aim to enhance communication and collaboration between these two groups (Chiheb et al., 2019a; Chiheb et al., 2019b). The use of interpretable ML models further facilitates this shared understanding. When domain experts and data analysts can easily interpret and comprehend the underlying decision logic of the model, it enables effective communication and collaboration (Lundberg & Lee, 2017). The transparency of interpretable models allows for meaningful discussions, validation of decisions, and alignment of perspectives, leading to a stronger shared understanding between domain experts and data analysts.

- **SO4**: The artifact creates a shared understanding between domain experts and data analysts.

The fifth solution objective of the artifact focuses on the learning from data element of data-driven decision-making. It aims to enable decision-makers to extract meaningful insights from data by identifying patterns, understanding them, and reflecting upon the extracted decision logic. This objective aligns with the learnability goal of the extracted decision logic, allowing decision-makers to acquire knowledge and insights from the model's decision logic. By leveraging these insights, decision-makers can assess risks, identify opportunities, and make well-informed decisions based on data-driven evidence. The extracted decision logic can also aid decision-makers as a decision support tool, further assisting them in their decision-making process.

- **SO5**: The artifact improves the overall decision-making process by learning from data.

Finally, the sixth solution objective of the artifacts is to address the gap identified in our literature review by incorporating domain knowledge at the DMN's decision requirements level and leveraging ML to extract the decision logic at DMN's decision logic level. Incorporating domain knowledge at the decision requirements level ensures that the decision-making process aligns with the specific needs and requirements of the domain. Simultaneously, the utilization of ML techniques at the decision logic level allows the artifact to extract the underlying decision logic, which may be implicitly contained in historical decision-making data (Bazhenova et al., 2016).

- **SO6**: The artifact incorporates domain knowledge at the decision requirements level and machine learning at the decision logic level.

4.2 Demonstration & Evaluation

These initial solution objectives are presented to practitioners through exploratory focus groups combined with individual questionnaires. The structured questions align with the evaluation

criteria defined earlier and are presented in Table 12.

Table 12: Evaluation questions

Criterion	ID	Question
Validity	Q1	The solution objectives accurately represent the problem and desired outcomes.
Completeness	Q2	The solution objectives cover all necessary aspects of the desired outcomes.
Generality	Q3	The solution objectives can be applied to various decision domains.

4.2.1 Results

All three data analysts (DA1, DA2, and DA3) found the solution objectives to be understandable, complete and generic. Their responses to the evaluation statements have been presented in Table 13.

Table 13: Evaluation 1 responses

ID	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Q1				DA1, DA3	DA2
Q2				DA1, DA3	DA2
Q3				DA1, DA2, DA3	

However, during the discussions, DA1 raised a concern that real-world data limitations could constrain the results of the project. To address this, it was suggested to add an additional solution objective (SO7) that would specify the identification of data limitations that restrict data-driven possibilities. This objective aims to explore and understand factors such as missing data or data quality issues that may hinder the utilization of data for decision-making purposes. By identifying and acknowledging these limitations, the project can provide valuable insights into the limitations of the available data. DA2 and DA3 agreed with DA1's proposal and indicated that revealing such data limitations is important to properly evaluate the results and could help improve data collection and thereby decision-making processes in the long term.

- **SO7**: The artifact reveals data limitations relevant for data-driven decision-making.

DA1 also emphasized that even with data limitations, the artifact should ensure a best effort is made with the data currently available. However, both DA2 and DA3 expressed their perspective that such best effort with the available data is already implicit in the proposed solution objectives. The objectives encompass the utilization of data for data-driven decision-making, inherently implying a commitment to maximizing the value derived from the available data. Therefore, while acknowledging the point made by DA1, it is concluded that the emphasis on making a best effort with the available data is already embedded within the scope of the solution objectives.

5. Phase 3: Artifact Design & Development

In this chapter, we describe the third phase of the artifact development. Based on the solution objectives determined in the previous chapter, an initial version of a DMN-based data-driven decision-making (DMN-D3M) framework is developed. We describe the artifact along with its design decisions and underlying motivation below.

5.1 Artifact description

The framework is described as hierarchical process framework, comprising three levels of abstraction: *phases*, *generic tasks* and *process instances* (Figure 8). At the highest level, the framework is organized into five *phases* (Figure 9). Each phase consists of several second-level *generic tasks*. The *process instance* level captures the actions, decisions, and outcomes of implementing the framework, based on the specified *phases* and *generic tasks*. The participation of two key roles, domain experts and data analysts, forms the foundation of the framework's execution (both defined in Table 14).

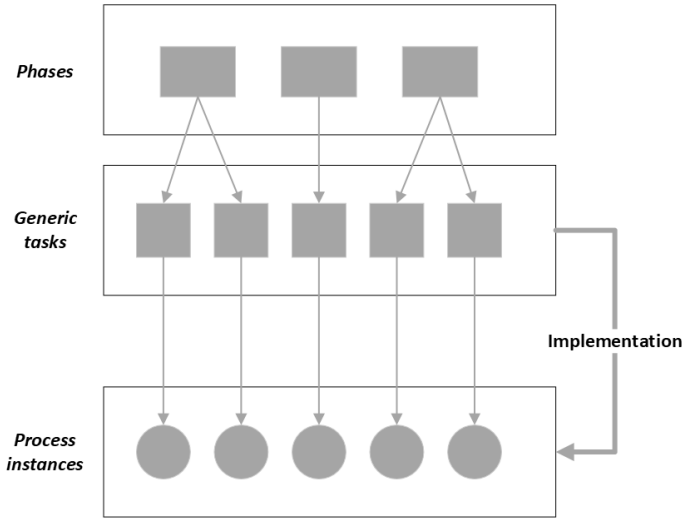


FIGURE 8: Framework structure

TABLE 14: Role definitions

Role	Definition
Data analyst	A professional who specializes in transforming data into insights by using a variety of analytical tools and techniques.
Domain expert	A professional who possesses a comprehensive understanding of the decision-making within a specific field, including its underlying processes, goals, and requirements.

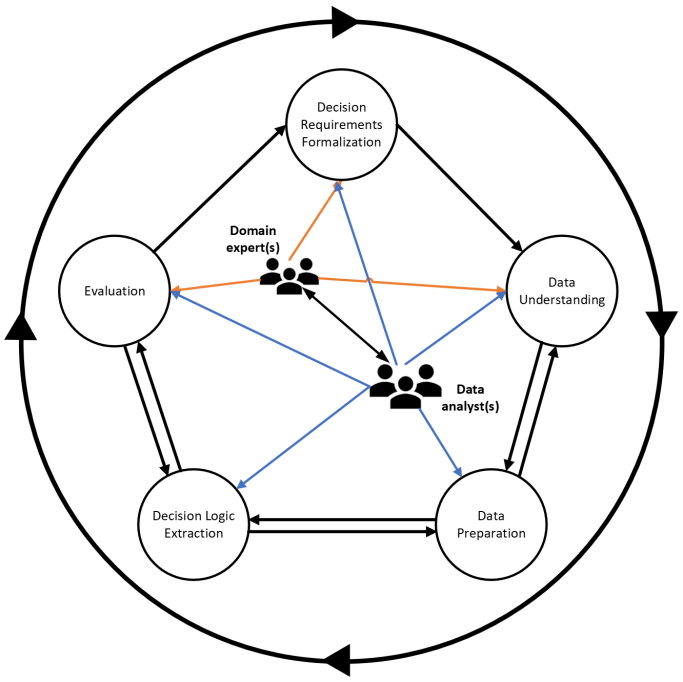


FIGURE 9: DMN-D3M Phases

The first version of the framework integrates DMN (OMG, 2015) with the widely adopted Cross Industry Standard Process for Data Mining (CRISP-DM; Wirth & Hipp, 2000) reference model. CRISP-DM offers a structured approach that guides organizations through the data mining process, making it a crucial element in the design of our framework. It is is considered the de-facto standard in ML (Schröer et al., 2021) and complementary to DMN, making the exploration of alternative options unnecessary. By systematically integrating DMN and CRISP-DM, we provide a decision-focused process framework for ML that prioritizes the involvement of domain knowledge throughout the process. Such frameworks are considered to be success factors in ML projects (Saltz et al., 2018) and support in understanding and managing the interactions within these complex projects (Wirth & Hipp, 2000).

At the top level, our framework consists of five phases: Decision Requirements Formalization, Data Understanding, Data Preparation, Decision Logic Extraction, and Evaluation. These phases align with the corresponding phases from CRISP-DM (Table 15), providing a comprehensive structure that ensures the formalization of domain knowledge and the generation of interpretable decision logic. Our framework leverages the decision requirements level of DMN to extend the business understanding phase of CRISP-DM, enhancing the formalization of decision requirements and capturing the decision-related aspects explicitly. Additionally, at the modeling phase of CRISP-DM, our framework incorporates the decision logic level of DMN, allowing for the modeling of decision logic using DMN's decision tables. The framework specifically targets scenarios where decision logic is not explicitly known in advance but is implicitly present in the data. In these cases, ML-based approaches are particularly useful for extracting this logic from the data (Bork et al., 2023).

To generate reliable and accessible insights, we recognize the importance of involving both data analysts (in blue) and domain experts (in orange). Therefore, we aim to create a shared

Table 15: Comparison of phases in CRISP-DM and DMN-D3M

CRISP-DM	**DMN-D3M**
Business Understanding	Decision Requirements Formalization
Data Understanding	Data Understanding
Data Preparation	Data Preparation
Modelling	Decision Logic Extraction
Evaluation	Evaluation
Deployment	-

understanding of both decision requirements and decision logic by their collaboration, ensuring that the data analysis is not isolated from the domain knowledge and vice versa. The framework emphasizes this continuous communication and collaboration by connecting the two roles in the middle. Furthermore, the outer circle symbolizes the cyclic nature of ML projects itself. The insights gained during the process often lead to new initiatives, making it an iterative and ongoing process.

We describe each *phase* and its corresponding *generic tasks* (summarized in Table 20 in Appendix A) along with its underlying motivation below. While these *generic tasks* provide a high-level description of the activities that should be carried out during each *phase*, they are not meant to be taken as an exhaustive or obligatory guide. Finally, the resulting artifact is documented through an ISDT (Table 21 in Appendix B).

5.1.1 Decision Requirements Formalization

The initial phase of the framework focuses on gaining an understanding of the decision-making process and its requirements. To achieve this, data analysts collaborate closely with domain experts to elicit their relevant domain knowledge, while utilizing DRD as a formalization technique to aggregate and represent this knowledge. Conceptual models, such as DRD, are particularly valuable in this context as they highlight relevant aspects that aid in understanding and communication among stakeholders, as highlighted by Mylopoulos (1992). The formalized decision-making knowledge is then presented back to domain experts to ensure its accuracy and completeness. This validation process not only confirms the correctness of the of the captured information, but also fosters a shared understanding of the decision-making process among stakeholders. In cases where additional contextual information is helpful, the formalization process can potentially be extended using additional conceptual modeling techniques such as BPMN or CMMN.

This initial phase of the framework builds upon the business understanding phase of CRISP-DM, focusing specifically on the decision-making process with the use of DMN's DRD. It marks the beginning of the project and forms the foundation for the execution of the remaining phases. Notably, Sharma & Osei-Bryson (2008) found that the real-world implementation of this phase in CRISP-DM is often performed in a rather unstructured and ad-hoc manner due to the absence of appropriate tools and techniques. To address this issue, our framework incorporates the DRD as an essential technique. This suggestion to use DRD as part of the development of business understanding draws inspiration from the work of Chiheb et al. (2019a and 2019b) who proposed

two conceptual models that utilize DRD to bridge the communication gap between domain experts and data analysts. However, these models have not yet been empirically validated in real-world scenarios.

5.1.2 Data Understanding

The data understanding phase starts with the initial data collection and proceeds with activities to gain an understanding of the data. However, it is crucial to recognize that data analysts cannot accomplish these tasks in isolation. While data analysts lead this phase, it requires collaboration and communication with domain experts. The active involvement of domain experts is essential to provide valuable context, domain-specific insights, and interpretations that contribute to a better understanding of the data.

By explicitly involving domain experts and emphasizing their integration within the data understanding phase, our approach extends the standard CRISP-DM framework to create a more holistic and mutually beneficial process that leverages the collective expertise of both data analysts and domain experts. Furthermore, the data understanding phase provides an opportunity for data analysts to assess the data quality and identify potential discrepancies between the available data and the required data specified in the DRD.

5.1.3 Data Preparation

The data preparation phase is a crucial step in the data analysis process and follows the methodology outlined in CRISP-DM. During this phase, data analysts perform all the necessary preprocessing steps to construct the final dataset, which should then be ready for the subsequent phases of the framework. This involves data cleaning, such as handling missing values, addressing outliers, and resolving inconsistencies. Additionally, if the data originates from different sources, data merging may be necessary to create a unified dataset. Furthermore, data analysts may build upon the insights generated from domain knowledge (DRD) to select and create relevant features that are considered relevant for the decision-making. This feature engineering step may involve transforming, combining, or creating new features that enhance the dataset's predictive power. It is important to note that there is a close link between data preparation and the previous phase, as raw data typically needs to be preprocessed before insights can be generated.

5.1.4 Decision Logic Extraction

The decision logic extraction phase involves selecting and applying algorithms to extract decision rules from the data. While the CRISP-DM modelling phase allows for the selection of any modeling technique, our framework specifically focuses on rule mining algorithms, or models that can be translated into decision rules. In this way, we can present the decision logic as decision tables, which has shown to offer several benefits, such as interpretability (e.g., Huysmans et al., 2011; Martens et al., 2007), maintainability (Lima et al., 2009) and learnability (Mues et al.,

2005). These ML-based decision tables align with the decision logic level of DMN and integrate with the DRD specified earlier.

To train a model, it is crucial to split the data into a train-test set, and potentially further divide the train set in a training and validation set to optmize the model's hyperparameters. This two-step approach is essential to prevent both overfitting and data leakage. Overfitting happens when the trained model is overly complex and fits the training data too closely, which results in over-optimistic performance on the training data, but poor generalizatibility to unseen data (e.g., test set). Data leakage occurs when information from the test data is used to train the model (e.g., hyperparameter selection), inflating the test performance.

It is important to recognize that the quality and relevance of the extracted decision rules is closely linked to the data preparation carried out in earlier phases. As such, the insights generated from the decision rules may necessitate revisiting the preprocessing steps to ensure that the decision rules reflect the underlying patterns present in the data as best as possible.

5.1.5 Evaluation

The evaluation phase plays a crucial role in assessing the performance and validity of the decision logic extracted from the data. It involves presenting one or more decision tables to domain experts, allowing for collaborative evaluation and obtaining their feedback. The purpose of this evaluation is twofold: to collaboratively assess the performance and meaningfulness of the extracted decision logic, and to reflect upon the insights generated through the analysis.

During this collaborative evaluation, domain experts have the opportunity to reflect on the decision logic extracted from the analysis and evaluate its alignment with their domain knowledge. This validation process is crucial as it ensures that the decision logic is not only technically accurate but also resonates with the expertise and expectations of domain experts. In an optimal scenario, the decision logic can be validated by domain experts and immediately incorporated into the decision-making process without a specific deployment activity, distinguishing our framework from the CRISP-DM framework.

Moreover, the evaluation phase also allows for the identification of potential mismatches between the derived decision logic and the modelled decision-making process through domain knowledge. For example, the DRD may represent an idealized scenario, but the analysis exposes shortcuts or deviations that occur in real-world decision-making. Additionally, this phase enables experts to reflect on the importance of variables in decision-making, leading to the discovery that certain variables previously considered crucial may have less influence than expected, or uncovering previously overlooked important factors. This reflective process enhances the overall understanding of the decision-making process and encourages experts to critically assess and refine their existing domain knowledge.

Furthermore, the evaluation phase may reveal that the available data is insufficient to capture relevant decision-making patterns effectively. Such findings represent crucial insights for future data-driven projects, emphasizing the need for specific data improvements. Still, it is important to acknowledge that this phase provides insights that may necessitate revisiting the decision logic

extraction, emphasizing the close connection between the evaluation phase and the previous phases of the framework.

5.2 Demonstration & Evaluation

This initial version of the artifact is then presented and evaluated through an exploratory focus group. The structured questions again align with the specified evaluation criteria (Table 16).

Table 16: Evaluation questions

Criterion	ID	Question
Understandability	Q1	The DMN-D3M framework is presented in a clear and easy-to-understand manner.
	Q2	The Decision Requirements level is clear and understandable.
	Q3	The Decision Logic level is clear and understandable.
Completeness	Q4	The DMN-D3M framework covers all necessary aspects of data-driven decision-making.
Efficacy	Q5	Relevant decision-making knowledge is formalized through a Decision Requirements Diagram.
	Q6	Interpretable decision logic is extracted from the data.
	Q7	The framework suggests improvements for the overall decision-making process by learning from data.
Generality	Q8	The concepts and activities in the DMN-D3M framework are generalizable to other decision-making scenarios.

5.2.1 Results

In the second focus group, the data analysts agreed upon most aspects of the framework, including its understandability, efficacy and generality (Table 17). After a short presentation, the group found the framework to be clear and easily understandable, and both DMN levels were also clear to them. They also agreed that the DRD could be used to formalize relevant decision-making knowledge and that decision tables were interpretable representations of the extracted decision logic. Additionally, the group believed that executing this framework could improve the overall decision-making process by learning from the available data. Although DA2 initially disagreed with question 7, the response was revised to agree after receiving clarification from the group. The group as a whole also confirmed the generality of the framework.

Table 17: Evaluation 2 responses

ID	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Q1				DA1, DA2, DA3	
Q2					DA1, DA2, DA3
Q3					DA1, DA2, DA3
Q4	DA1		DA3	DA2, ~~DA3~~	
Q5				DA1	DA2, DA3
Q6				DA1, DA2	DA3
Q7		~~DA2~~		DA2	DA1, DA3
Q8				DA1	DA2, DA3

However, the data analysts did not form a shared opinion on the completeness of the framework during the focus group. While DA2 and DA3 confirmed its completeness, DA1 strongly disagreed and argued for the inclusion of guidelines to handle missing data. Specifically, DA1 suggested making assumptions on the decision logic if data is missing to generate artificial data based on the assumed decision rules (e.g., through simulation). This suggestion is similar to the approach by Bork et al. (2023), although they were unaware of the paper at that time. DA1 emphasized that this is especially relevant in the civil engineering field, where missing data is common and artificial data generation could potentially address this. DA3 initially agreed with the completeness, but started to doubt during the discussion and ultimately revised to neutral.

However, the suggestion to improve the completeness by incorporating guidelines for handling missing data was not implemented. The framework focuses on extracting the decision logic from historical data and making assumptions on the decision logic to generate artificial data does not to contribute to this goal. The generated data based on assumptions would only confirm the existing decision logic, offering no new insights. If data limitations exist, SO7 states that they should be revealed. Additionally, the proposed addition assumes incomplete data, which may not be applicable to all decision domains. Moreover, adding artificially generated data may reduce the interpretability of the results, which goes against one of the main objectives of the framework. Therefore, considering the objectives established in the first focus group, guidelines for dealing with missing data were not included in the framework. Upon explanation, DA1 agreed with this decision.

6. Phase 4: Artifact Demonstration & Evaluation

In this chapter we first demonstrate the framework through a process instance in a real-world setting. Thereafter we perform the ex-post evaluation to assess the artifact.

6.1 Artifact Demonstration

As described earlier, a rail maintenance setting at RHDHV has been selected to demonstrate the artifact. In this setting, a condition-based maintenance strategy has been implemented, which involves yearly inspections to monitor the rail's condition and make preventative maintenance decisions. To demonstrate the framework, we focus on the maintenance decisions for curves, as these are some of the most critical sections where the risk of derailment increases with curvature. Furthermore, additional measurements are taken in curves, which provides us with more data compared to straight sections. Within the possibilities of the available budget, the safety of the rail network is the top priority. Nevertheless, RHDHV acknowledges the need to enhance its understanding of how rail experts make maintenance decisions and aims to leverage the historical data to improve their decision-making process.

6.1.1 Decision Requirements Formalization

The first phase of our framework involves collaborating with domain experts to formalize the decision-making process using a DRD, as shown in Figure 10. The maintenance decision is central to this process and forms the basis for a multi-year investment plan (MIP) that outlines the anticipated maintenance activities for the upcoming four years. In this maintenance plan, the upcoming year is fixed, while the plan for the next three years is flexible and can be adjusted based on the level of degradation in the upcoming years. This approach has two benefits according to the decision-makers. First, it ensures that maintenance decisions are based on the current condition of the rail system, rather than predetermined plans. Second, it enables decision-makers to spread the anticipated maintenance based on their estimated costs, ensuring that the total costs remain stable. As observed in the DRD, the maintenance decision depends on three high-level inputs: condition grades, measurements and contextual factors. Each of these input groups is described below.

First and foremost, condition grades are collected annually for 4 components: rail, sleeper, ballast and dimensions (Figure 11). While there are several criteria for each component that domain experts evaluate, the final grade is determined based on the worst part of the section, relying on expert judgment. The four individual grades are also combined into an overall grade, with pre-determined weights assigned to each component (2, 4, 1, and 2 respectively).

In addition to grading these four components, three additional measurements are taken for curved sections: abrasion, cant, and track gauge. Abrasion is wear that results from the contact between the train's wheels and the rail. Cant refers to the angle between the left and right rail

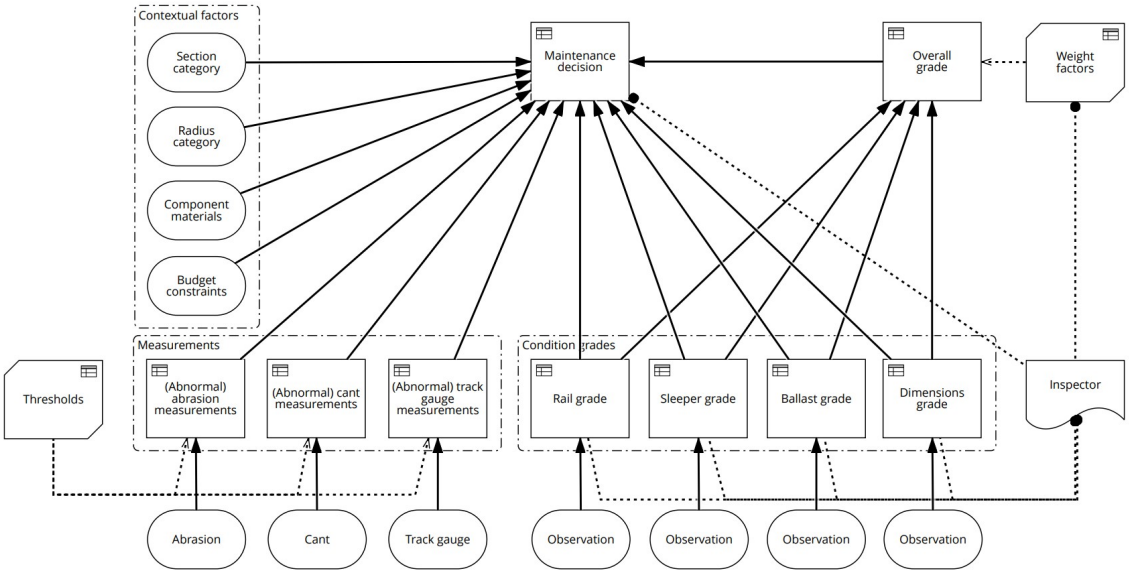


FIGURE 10: DRD rail maintenance decisions

(Figure 12), which is intentional in curvatures to alleviate forces acting on the track. However, an excessively varying cant along the track can lead to track twist. Track gauge refers to the horizontal difference between the left and right rail (also shown in Figure 12). For all measurements, a cautionary and safety threshold is in place to identify abnormal values. When a measurement exceeds the cautionary threshold, caution is advised, but the section can still operate. However, if a measurement crosses the safety threshold, maintenance should be planned accordingly. At RHDHV, the goal is to maintain all measurements within both thresholds to prevent any potential safety hazards.

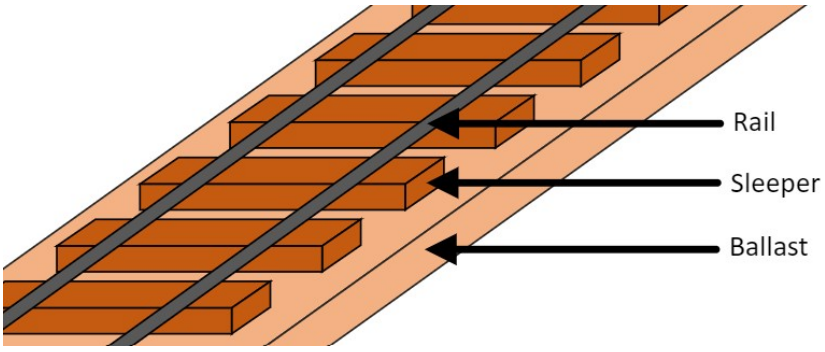


FIGURE 11: Grade components

Finally, several contextual factors are taken into account when making maintenance decisions. For instance, rail sections are prioritized based on their usage (A, B, or C). And curved sections are also assigned to either category 1 or 2, depending on the radius of the curvature, which affects the amount of force acting on the track. Each component's material is another factor that is considered, as some materials are more vulnerable than others to degradation and wear. Additionally, the budget is a key consideration, as a fixed budget is available each year, and decision-makers must optimize its use to ensure that the costs remain stable. From that perspective, the current maintenance strategy can be also be considered budget-driven, as it prioritizes maintenance activities based on the available budget.

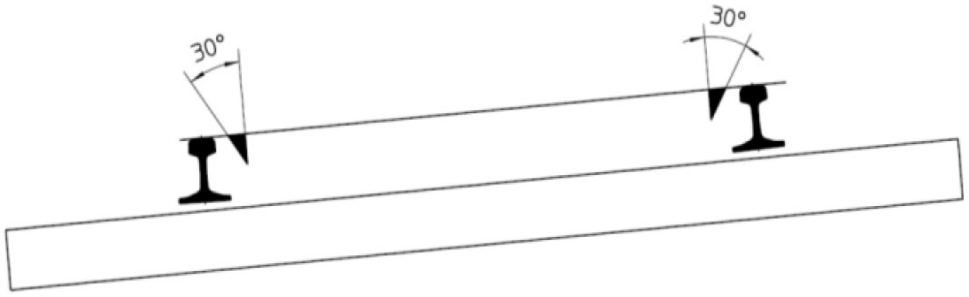


FIGURE 12: Cant and track gauge example

6.1.2 Data understanding

In the second phase of our framework, we collect and examine the relevant data from the past 12 years of inspection records. To provide an initial overview of the data, we present a visualization of the overall grade distributions over the years in Figure 13, as well as visualizations of the individual component's grades in Figures 19 to 22 (Appendix C). It should be noted though that the component weights (to combine the four component grades) were different before 2019, but have been updated with the latest component weights to ensure consistency across all years. Also, the sleeper grade used to be split based on the type of rail mount on the sleeper (direct or indirect). In such cases, we kept the lowest grade of the two, as discussed with the domain experts.

The visualizations reveal several noteworthy trends that provide insight into the maintenance of the railway network. First, the number of curves graded in the network has remained consistent over the years, indicating the completeness of our data. Second, there is a clear downward trend in the proportion of curves with low condition grades. In 2011, the average condition grade was 6.0, while in 2022, it had improved to 7.2. This suggests that the network's condition has gradually improved over time, a trend that domain experts attribute to effective maintenance decisions made within the available budget. However, the 2011 data was excluded from our analysis due to a significant increase in the grades between 2011 and 2012 that could not be explained by maintenance events. Upon further investigation, domain experts confirmed that the grading procedure had been slightly modified after the first year of grading. Third, a shift in grading is observed in the high grades from 2018 to 2019, where higher grades (e.g., 9) seem to have been replaced with slightly lower grades (e.g., 8). This change was unexplainable by domain experts.

The measurement data, such as abrasion (Figure 14a) and cant (Figure 14b), also show a decreasing trend in abnormal values over time, indicating an improvement in the rail network. However, this trend is not apparent in track gauge (Figure 14c), where abnormal values have always been rare. It's worth noting that the number of measurements in these visualizations is approximately a factor 250 higher than in the condition grades, as measurements are taken every 10 sleepers, while condition grades are assigned per section.

In terms of contextual data, the sections are categorized as approximately 25% category A, 41% category B, and 34% category C. The distribution of sleeper materials over the years is shown

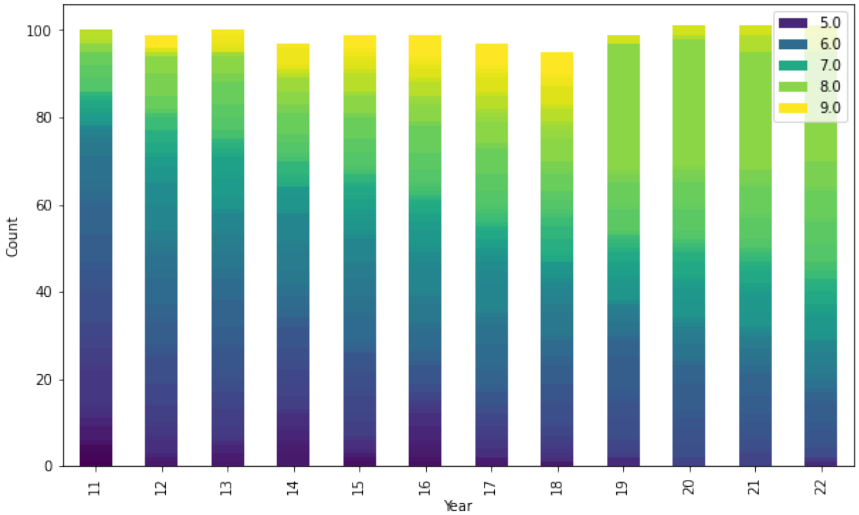
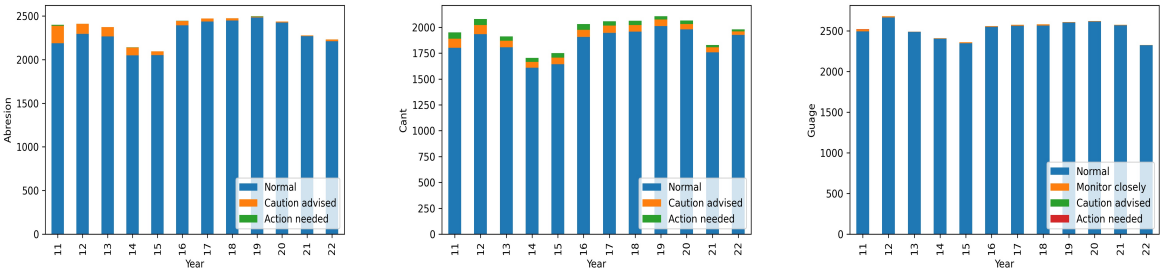


FIGURE 13: Condition Grades by Year



(A) Abrasion (B) Track twist (C) Gauge

FIGURE 14: Measurements by Year

in Figure 15, indicating a gradual replacement of wooden sleepers with concrete ones, which generally have a longer lifespan. However, the curve categories are incomplete, with 65% of all sections missing. In the available data, 20 sections fall under category 1 and 17 sections fall under category 2.

The final dataset comprises 143 maintenance events, which we plotted against against the overall condition grade (Figure 16) as recommended by the rail experts, who rely mainly on the grading system for their maintenance decisions. Indeed, maintenance activities are predominantly scheduled when the section's condition grade is relatively low, indicating the significance of condition grades in decision-making. However, we identified several outliers that were discussed with the domain experts. During this discussion, we discovered and removed one incorrectly registered maintenance event and one incidental event. Two outliers were explainable through individual component's grades. The remaining outliers left the experts puzzled, suspecting that higher grades were assigned to the section than it's weakest part (inconsistent grading).

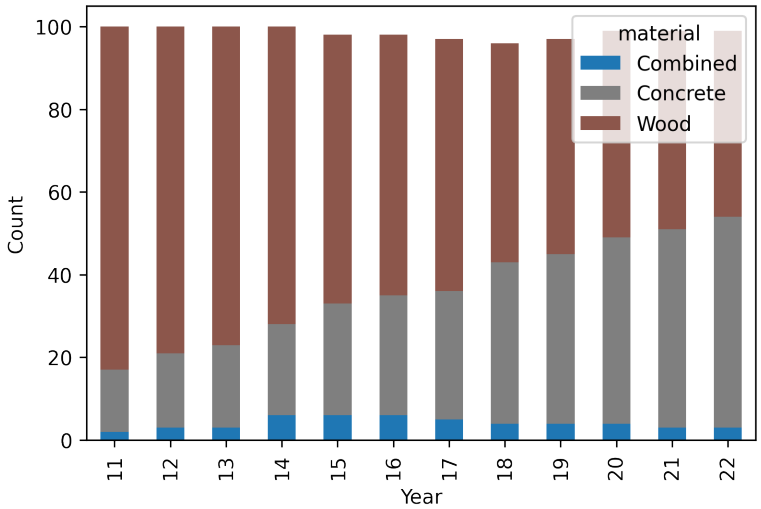


FIGURE 15: Material Distribution by Year

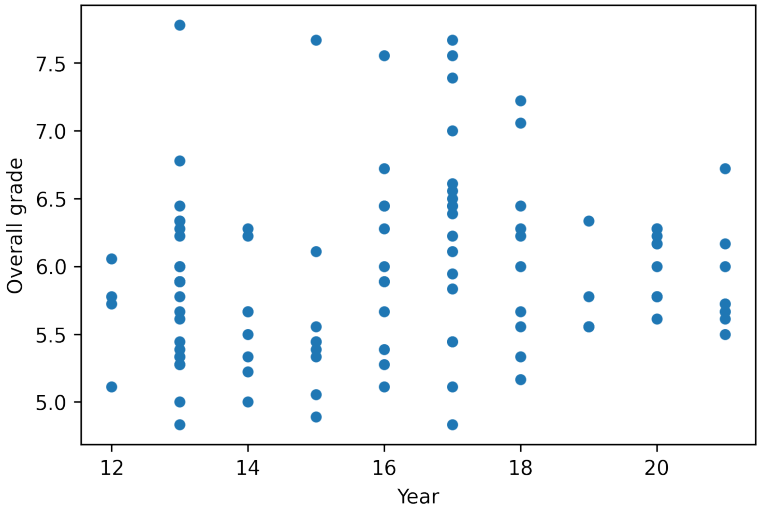


FIGURE 16: Maintenance events per year against the overall condition grade

6.1.3 Data Preparation

To extract meaningful insights from the raw data and to prepare the final dataset for rule extraction, we utilized the Pandas and NumPy libraries in Python to perform various preprocessing steps. The data was originally stored across multiple Excel files and the management system with varying formats used over the years. To address this, we began by extracting all relevant inspection data, and then standardizing and cleaning it. We then (inner) merged the inspection data by matching the year and object combination. However, we encountered several challenges during this process.

1. The first challenge we encountered was that the measurements (abrasion, cant and gauge) are taken at every 10 sleepers, which results in a varying number of measurements per object as sections have varying lengths. To address this issue, we transformed the measurements based on the corresponding thresholds of the specific measurement type, ensuring

the same number of variables per section, independent of its length. Additionally, we included both the mean and maximum measurement per section to mimic how inspectors analyze the data (e.g., not only look at abnormal values, but also skim through the data to get a feeling of the measurements). The mean value provides insight into the overall condition of the specific measurement, while the maximum value shows how close a section may be to crossing a threshold. For cant measurements, we transformed this into track twist based on the difference in cant for every 20 sleepers (12 meters).

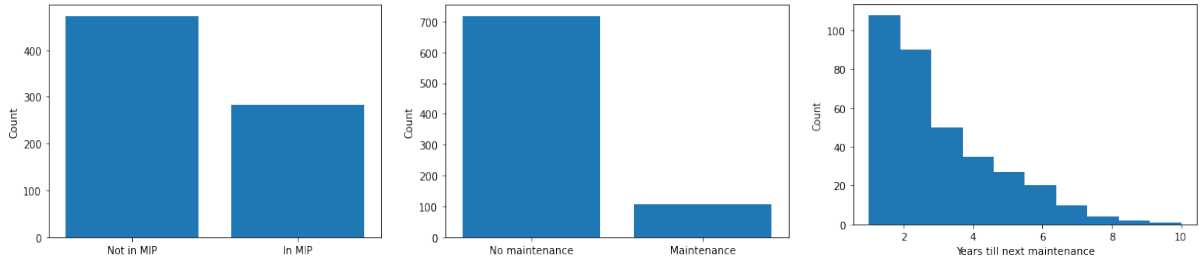
2. The second challenge was that sometimes multiple measurements are taken at the same section in the same year, especially when abnormal measurements are found. In such cases, the measurements are reproduced to confirm or disprove the abnormal values, since these manual measurements are prone to errors, and inspectors may question their reliability. We only retained the latest measurements in our dataset, which are considered most reliable.
3. The third challenge was that some sections are sometimes divided into sub-sections when they are separated by an obstacle (e.g., a crossing), resulting in multiple objects in the dataset (e.g., 700-1 and 700-2) that refer to the same section (e.g., 700). This is particularly challenging in combination with the previous issue, as some sub-sections may be remeasured, while others are not.
4. The fourth challenge was missing values in the datasets. For variables with few missing values, we removed rows with missing data. However, we faced challenges with several variables that had many missing values: all component materials and the curve category. Given the importance of the sleeper material, we leveraged domain knowledge to impute the missing sleeper materials and completed the data. Other components were excluded from the analysis as they were deemed of lesser importance by domain experts. For curve categories, obtaining this data would require extensive measurements, and we decided to create two separate datasets, one including curve categories and another without them.

Additionally, with a fixed yearly budget spent on maintenance and a continuously changing distribution in the network's condition, it is important to determine the criticality of sections relative to each other. To achieve this, we ranked each section based on its overall condition grade and minimum component's grade for every year. This enables us to assess each section's importance relative to others, regardless of its absolute grades.

After this initial data preparation phase, we extracted the maintenance data from the management system and (outer) merged it with the inspection data. We then calculated the number of years until the next maintenance event using this data. Since maintenance is always planned one year ahead, we removed rows with exactly zero years until the next maintenance event. To ensure the correctness and completeness of the dataset, we evaluated the maintenance data against the condition data and removed incorrect maintenance events and added missing maintenance events (already included in Figure 16). We also slightly adjusted some maintenance dates to match the observations made in the data. Furthermore, we excluded all rail sections that were or had been out of use based on the information in the management system.

In the Decision Logic Extraction phase, we explored three scenarios to extract decision logic: (1) predicting whether a rail section should be included in the MIP covering the next four years, (2) predicting whether maintenance should be planned in the upcoming year, and (3)

predicting the exact number of years until the next maintenance event. Figures 17a, 17b, 17c present the target variable distribution and dataset size for each scenario respectively. In order to incorporate instances where the exact number of years till the next maintenance event is unknown (as it has not yet occurred) in scenarios (1) and (2), we included data from sections where no maintenance was performed within 4 years as negative examples for scenario (1), and within 1 year for scenario (2). To encode the categorical features, we used one-hot-encoding, which transforms categorical variables into binary representations, creating dummy variables for each unique category.



(A) Scenario 1 (B) Scenario 2 (C) Scenario 3

FIGURE 17: Prediction scenarios

6.1.4 Decision Logic Extraction

To facilitate decision logic extraction, the dataset is split into a train (80%) and test (20%) set using using stratified sampling with the target variable as strata. We then used 10-fold cross validation within the training set to determine the model's optimal set of hyperparameters. This approach using cross validation rather than a separate validation set is suitable for our relatively small dataset (with less than 1000 instances) and simple algorithms (that do not require extensive compute).

In order to extract the decision rules from the data, we selected two tree-based algorithms: C4.5 and an optimized version of CART (Classification And Regression Tree; Steinberg, 2009). C4.5 constructs decision trees that can have more than two child nodes, while CART constructs binary decision trees. We also made an attempt to incorporate the algorithm proposed by Simić et al. (2020), but the provided description was insufficient for full replication. Unfortunately, its performance suffered when we made assumptions about how the algorithm worked, and it did not align with the results reported in their paper. This discrepancy suggests that either our assumptions or their reported results may be incorrect.

Due to the imbalanced target variable distribution in all three prediction scenarios, the accuracy metric is not suitable to measure the model's performance. This is because a model could achieve a seemingly good accuracy by simply predicting the majority class for all instances, while misclassifying all minority classes. Precision and recall are more appropriate metrics for evaluating model performance in such scenarios. Precision measures the model's ability to avoid false positives, while recall measures its ability to identify all positive instances. However, a higher recall typically leads to a lower precision, so the F1-score represents a useful single score metric that summarizes the model's performance, which is also used for hyperparameter optimization.

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{F1-Score} = 2 * \frac{Recall*Precision}{Recall+Precision}$$

For each scenario, we followed the procedure as described above. However, we obtained unsatisfactory performance measure in scenarios (2) and (3), which we have included in Appendix D. Additionally, including curve categories resulted in decreased performance due to the reduced amount of available data, and did not offer significant added value. As a result, we will focus on the results of scenario (1) with the curve category excluded in the subsequent section.

Hyperparameter optimization was feasible for CART, which was implemented using the scikit-learn library (*DecisionTreeClassifier*), but not for C4.5, which was implemented with the Chefboost library. For CART, we set the *class weight* to 'balanced' to address the imbalanced data during training by adjusting weights inversely proportional to the class frequencies. The optimized hyperparameters were the *max depth*, representing the maximum depth of the tree and *max leaf nodes*, indicating the maximum number of leaf nodes. These two hyperparameters where selected as they represent the maximum width (length of each decision rule) and length (the number of decision rules) of the resulting decision table. Other hyperparameters were left at the default setting. The results of the hyperparameter tuning are presented in Figure 18.

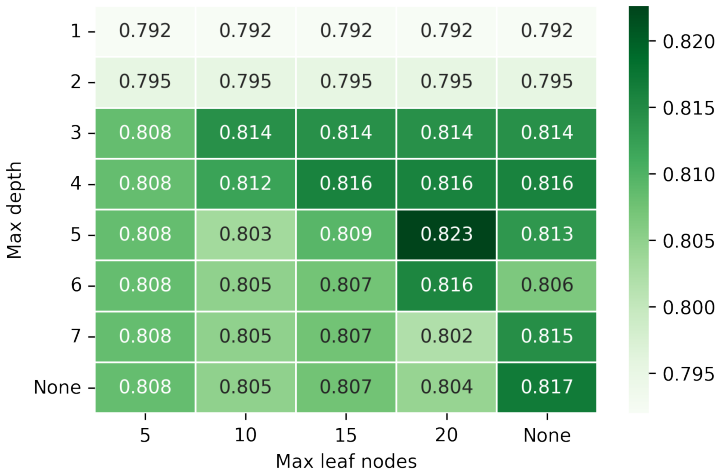


FIGURE 18: CART hyperparameter optimization

Through this hyperparameter optimization we found that the optimal performance (F1-score of 0.823) is achieved by setting the *max depth* parameter to 5 and the *max leaf nodes* parameter to 20. Surprisingly, further increasing the tree depth beyond 1 had only marginal impact on improving the F1-score, while increasing the complexity of the rules at each depth level. Even with minimal complexity, the alternative tree configuration performed similarly, scoring only 0.03 lower than the optimum the F1-score. Still, the results of the optimal tree generalized well to the test set, with an F1-score of 0.80. In contrast, C4.5's hyperparameters could not be optimized, resulting in a training F1-score of 0.876 that did not generalize to the test F1-score (0.762), which indicates overfitting on the training data. The resulting classification matrix on the test set, feature importances and resulting decision table are included in Appendix E.

6.1.5 Evaluation

While the extracted decision rules demonstrate reasonable performance, their practical applicability is currently limited. Domain experts confirmed that the model's additional complexity is not beneficial and is likely capturing noise in the data rather than capturing the reality of decision-making (overfitting). Instead, a simpler decision tree with minimal complexity, which utilizes the overall condition grade (threshold = 6.81) to split the data, aligns better with the domain experts' expectations as they consider the overall grade to be the primary driver of maintenance decisions. However, the concern regarding the reliability of both models persists. For example, the currently determined cutoff grade determined on historical data may not be reliable in the future due to the observed improvements in the network's grade composition over time.

Moreover, the reliability of the underlying condition grades is questionable due to the underlying expert judgment and potential (unconscious) changes in the grading policy over time. During the discussion of the results, the domain experts were surprised to discover that an important contextual factor such as section category was not considered to be important by the models. This realization led them to speculate that the contextual factors (section category, radius category and component material) might be implicitly incorporated during the grading rather than being independently taken into consideration during maintenance decision-making. For example, a section categorized as A may receive a grade of 6, while an identical condition in a category C section may receive a grade of 7. Consequently, this new insight deviated from our initially formalized and validated DRD, and further highlights the inherent unreliability of subjective grades.

Furthermore, an important aspect that is not included in the current analysis is the cost consideration. Although we identified that maintenance decisions are driven by budgetary constraints, it is infeasible to estimate maintenance costs due to the limitations of the available data. While maintenance decisions are recorded at the section-level, the reality is that maintenance decisions are made at the element-level. For example, only a subset of sleepers is typically replaced, not the entire section. Accurate cost estimations rely on detailed information about the extent of replacement required, which is currently unattainable as condition grades are provided per section and measurements are taken for every 10 sleepers. Consequently, our condition data is incomplete, and cost considerations cannot be incorporated. In practice, decision-makers rely on more detailed observations beyond what is captured in the data, allowing them to make more informed decisions.

6.2 Artifact Evaluation

After demonstrating the framework, we evaluated it with the domain experts and data analysts.

6.2.1 Data Analysts

The results of the data analysts are presented in Table 18

Table 18: Evaluation 3: Data analyst responses

ID	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Q1				DA1, DA2	DA3
Q2			DA1, DA2		DA3
Q3			DA2, ~~DA3~~		DA1, DA3
Q4			DA1		DA2, DA3
Q5				DA1	DA2, DA3
Q6					DA1, DA2, DA3
Q7			DA1, DA2	~~DA2~~, DA3	
Q8				DA1, DA2, DA3	

All three domain experts confirmed that they found the framework understandable. However, when it came to the specific DMN levels, data analysts held differing opinions on its understandability, which was contrary to their unanimous agreement in the ex-ante evaluation. This difference in opinion can be attributed to the fact that the first focus group discussed a simple mortgage scenario as an example, while the real-world implementation involved more complexity with a larger number of variables, including some that were less intuitive.

As a result, it became apparent during the evaluation that additional clarification was necessary, especially for those who were unfamiliar with the specific data. For example, in case of the DRD, concerns were raised about "unclarity about what a data point specifically means" and the need "to have more clarification to understand what is meant" (DA2). On the other hand, DA3 reported a positive experience, stating a "good and clear understanding of the diagram." Similarly, the decision tables were considered difficult to understand without any additional explanations. One data analyst remarked, "Just looking at the result, I couldn't get it. But with the explanation of what it means, how it works, I can understand it 100%" (DA3). DA1 agreed with this viewpoint, saying it is "clear with explanation, but not understandable without." These observations match other concerns stating that they "don't know the actual values itself and what they mean" (DA3) and they do not have "the deeper knowledge of what the data/numbers mean" (DA2). It is thus essential to provide comprehensive explanations that clarify the workings of DMN and the specific variables. Without such guidance, data analysts may face difficulties in understanding both DMN levels. However, given sufficient explanation, all three domain experts agreed on the framework's efficacy in capturing relevant domain knowledge through a DRD and extracting interpretable decision logic from the data.

In terms of the framework's completeness, all data analysts showed improvement in their perceptions compared to the first focus group. DA1's rating changed from strongly disagree to neutral, DA2's from agree to strongly agree, and DA3's from neutral to strongly agree. However, while DA1 indicated that the framework "touches upon all relevant aspects", concerns were still raised about its ability to handle missing data." Although revealing data limitations is valuable, and specific limitations are uncovered, practical recommendations on how to collect better data are absent" (DA1). On the other hand, DA2 disagreed with this perspective,

indicating that "changing the data collection process falls outside the framework's scope" and argued that "all relevant phases were well represented". Similarly, they did not reach a consensus regarding its ability to suggest improvements for the overall decision-making process by learning from data. Initially, DA2 and DA3 agreed on this, but DA1 emphasized that, despite generating insights and revealing data limitations, there were no practical suggestions provided to enhance the decision-making process. Taking into account this perspective, DA2 began to doubt the framework's capability to address data problems and switched to neutral.

Finally, all data analysts agreed on the framework's generality, indicating that it is a "good and clear approach, applicable to other areas" (DA1). However, DA2 noted "there are bound to be edge cases where it is not applicable". For example, "it will work only for operational decision-making, not necessarily tactical or strategic decision-making" (DA2). This observation indeed aligns with DMN, which is most suitable for operational decision-making. Additionally, DA3 added that "most likely it can be applied to other domains, but we cannot be 100% sure as some domains might pose difficulties, but in general yes".

6.2.2 Domain experts

The results of the domain experts are presented in Table 19.

Table 19: Evaluation 3: Domain expert responses

ID	**Strongly Disagree**	**Disagree**	**Neutral**	**Agree**	**Strongly Agree**
Q1				DE1, DE2, DE3	
Q2				DE1	DE2, DE3
Q3			DE1, DE3	DE2	
Q4				DE1, DE2, DE3	
Q5				DE1, DE3	DE2
Q6			DE2, DE3		DE1
Q7				DE1, DE2	DE3
Q8				DE2, DE3	DE1

The domain experts also confirmed that they found the framework understandable, although "it requires a bit of explanation, which was provided" (DE2). They particularly appreciated the DRD for "its ability to visually represent the relevant decision aspects in a clear and organized manner" (DE3). "The DRD served as a helpful tool that facilitated discussions and provided insights into the decision-making process" (DE3). The domain experts found such visualizations to be "more effective than relying solely on stories or bullet points" (DE3). By formalizing and validating knowledge through this visual representation, the domain experts also "felt more confident in the data analyst's understanding of the underlying decision-making process" (DE1), leading to increased trust in the project.

However, some concerns were raised by DE1 and DE3 regarding the understandability of the decision logic level. They felt that significant explanations were needed to comprehend how to interpret the decision table. They noted that "as the decision table became more complex with additional rows and columns, its interpretability diminished" (DE2). The decision table of the optimal model was considered too elaborate and a simpler version would have been preferred. This perceived lack of understandability also affected the domain experts' opinion on the framework's ability to generate interpretable decision logic. During the presentation of the results, the domain experts found the model's feature importances more informative than the decision table itself, and the full feature importances could not be derived from the decision table alone. They would "prefer visualizations rather than the numbers in the decision table" (DE3). "The step-by-step explanation towards and of the decision table makes it somewhat understandable, but in itself it is not" (DE3).

Furthermore, all domain experts agreed on the completeness of the framework. DE3 specifically highlighted the value of formalizing and validating the DRD, "ensuring that all crucial components were considered in the analysis" and thereby ensuring the completeness of the analysis itself. The domain experts also recognized that the framework effectively formalized relevant decision-making knowledge through the DRD. However, they emphasized that "the quality of this formalization depended on the data analyst's capability to accurately and understandably model it" (DE2). For instance, the three high-level boxes in Figure 10 were appreciated, but this was a modeling decision by the data analyst. The usefulness of the DRD therefore "depends on the abilities of the data analyst to model and present such diagram in an understandable way" (DE2).

Regarding the framework's efficacy in suggesting improvements for decision-making through learning from data, the domain experts noted that "the insights generated from historical data provided a factual basis for discussions and potential future improvements" (DE3). "This contributed to substantiating and clarifying discussions around data-driven decision-making" (DE3). However, the domain experts also noted that the framework "didn't offer concrete suggestions that could be directly implemented" (DE2). Finally, all domain experts agreed on the framework's generality, with DE3 expressing trust in its success in other domains: "This model is really going to work". However, DE1 cautioned that the applicability and efficacy of the framework is heavily reliant on the availability of historical data and expertise of domain experts and data analysts.

6.2.3 Comparison

When comparing the opinions of the domain experts and data analysts regarding the framework, several similarities and differences can be observed.

Both roles agreed that the framework was generally understandable. However, data analysts had differing opinions on the understandability of the specific DMN levels. While domain experts could intuitively understand the content and relationships in the decision requirements level, data analysts struggled and required additional explanations. This could be due to the domain experts' familiarity with the displayed information, which was not shared by the data analysts. Regarding the decision table at the decision logic level, both roles faced challenges

in understanding it. However, with additional explanations, data analysts found the decision table more interpretable, while the difficulties for domain experts remained.

In terms of the framework's ability to suggest improvements for decision-making through learning from data, both data analysts and domain experts had doubts about its capability to address data problems and provide practical suggestions. However, domain experts appreciated the insights generated from historical data, which helped substantiate and clarify discussions around data-driven decision-making. Although the framework didn't offer concrete suggestions that could be directly implemented, it provided a factual basis for potential future improvements.

Both data analysts and domain experts agreed on the framework's generality and its applicability to other areas. They recognized its potential to be adapted to different domains and decision-making contexts.

7. Discussion

In this chapter, we will discuss the extent to which our DS study has met the solution objectives defined for our framework in Phase 2: Solution Objectives.

- SO1: The artifact provides an understandable, complete and generic process framework for data-driven decision-making.

The feedback from domain experts and data analysts confirms that our framework has achieved a high level of understandability and generality. Users were able to quickly grasp the underlying concepts, principles, and components of the framework, indicating its potential for an efficient and effective application. Additionally, its potentially wide applicability remains unquestioned. However, there is some disagreement among data analysts regarding the completeness of the framework. While some data analysts acknowledge its completeness, others express concerns about the lack of concrete suggestions for improving the decision-making process, especially in scenarios with data limitations. Therefore, the framework largely meets SO1, but there are doubts about its completeness.

- SO2: The artifact describes how to formalize, aggregate and validate domain knowledge relevant for decision-making.

Both domain experts and data analysts validate the framework's capability to formalize, aggregate, and validate domain knowledge, thus meeting SO2. The DRD proved to be a valuable tool in visually representing relevant decision aspects and facilitating discussions. Domain experts particularly appreciated the clear and organized representation provided by the DRD, leveraging their domain knowledge to intuitively grasp the diagram's content. However, some data analysts found it more challenging to understand the relationships depicted in the diagram, highlighting the need for additional explanations for those less familiar with the domain.

- SO3: The artifact generates decision logic that is interpretable, maintainable, and learnable.

Contrary to our initial expectations and in contrast to previous studies (e.g., Lima et al., 2009; Mues et al., 2005), the framework does not fully meet SO3. Data analysts generally agreed that the decision logic generated by the framework is interpretable in the ex-ante evaluation. However, there is a division of opinions regarding its understandability in to the ex-post evaluation. The real-world demonstration introduced increased complexity, with larger decision tables and less intuitive variables. These findings align with the observations made by Hasić & Vanthienen (2019), who also noted the difficulties posed by complex decision tables. Domain experts found the decision tables even less understandable, which could be attributed to their relative unfamiliarity with this specific type of representation. It appears that they have a preference for visual representations, such as graphs, as opposed to tables filled with numerical data. Consequently, it became evident that the feasibility of a purely DMN-based approach depends on the complexity of the decision logic and the cognitive capabilities of individuals in dealing with such complexity (Bork et al., 2023). Despite these challenges, it is worth noting that the

interpretability of decision tables remained sufficient to invalidate the extracted decision rules, demonstrating that the framework partially meets SO3.

- SO4: The artifact creates a shared understanding between domain experts and data analysts.

The framework partially accomplishes SO4. The formalization and validation of domain knowledge through the DRD contributes to a shared understanding between domain experts and data analysts regarding underlying decision-making process. As a result, the domain experts feel more confident in the data analyst's understanding of the decision-making context and are ensured that all crucial components are considered in the analysis, leading to increased trust in the project. These findings align with the theoretical statements made by Chiheb et al. (2019a) and Chiheb et al. (2019b). Additionally, a shared understanding of the data is established through the presentation and discussion of insights during the data understanding phase. However, the limited interpretability of the decision tables limits the shared understanding in the final phase of the framework.

- SO5: The artifact improves the overall decision-making process by learning from data.

Domain experts confirm that the framework improves the overall decision-making process by learning from data. Executing the framework provides them with new insights from historical data and the extracted decision logic allows them to reflect on and refine their domain knowledge. However, the data analysts' initial agreement regarding the framework's potential to enhance decision-making through data learning shifted during the real-world demonstration. Although the framework generated new insights, revealed data limitations and contributed to the discussion around data-driven decision-making, it failed to provide practical suggestions for improving the decision-making process, leading to divided opinions among the data analysts. Hence, we can conclude that the framework partially fulfills SO5.

- SO6: The artifact incorporates domain knowledge at the decision requirements level and machine learning at the decision logic level.

The framework successfully incorporates domain knowledge at the decision requirements level and integrates ML at the decision logic level, meeting SO6. The formalization and validation of domain knowledge through the DRD enables the integration of domain knowledge into the ML project. However, in our demonstration, the ML-based decision tables were not representative for the actual decision logic, primarily due to data limitations.

- SO7: The artifact reveals data limitations relevant for data-driven decision-making.

Both domain experts and data analysts confirm the framework's ability to reveal data limitations relevant to data-driven decision-making. The process of decision requirements formalization, data understanding and decision logic extraction reveals the constraints and challenges posed by the currently available data. These insights stimulate critical reflection on data maturity and highlight the need for data improvements to transition towards data-driven decision-making. Therefore, the framework successfully meets SO7.

8. Conclusion and future work

In this concluding chapter, we will review our research objective, acknowledge the limitations of our study, and propose directions for future research.

8.1 Conclusion

This study aimed to address the challenges associated with the lack of methods for eliciting and validating domain knowledge in ML projects and the limited interpretability of black box models, which hinder the reliable incorporation of domain knowledge into data-driven decision-making. The research objective was to develop a novel framework that integrates domain knowledge and generates interpretable ML models to enhance the decision-making process.

To address these challenges, we investigated the application of DMN as a notation for modeling and incorporating domain knowledge and interpretability in ML models. By integrating DMN with the CRISP-DM reference model, a decision-focused framework was developed that emphasizes the involvement of domain knowledge throughout the ML model development process. This framework specifically targets scenarios where the decision logic is initially unknown and leverages the capabilities of ML to extract this logic from historical data.

The developed framework successfully addresses the first aspect of the problem statement by formalizing and validating domain knowledge using the DRD at the decision requirements level of DMN. The DRD serves as a visual and intuitive tool that facilitates discussions and generates insights into the decision-making process, fostering a shared understanding between domain experts and data analysts. Additionally, the DRD acts as a foundation for subsequent activities, such as data collection and feature engineering, thereby connecting the decision requirements to the decision logic level.

The second aspect of the problem statement is addressed by using decision tables at the decision logic level of DMN. However, the real-world demonstration revealed that in complex scenarios, the interpretability of decision tables may become a concern. Therefore, the feasibility of a purely DMN-based approach depends on the complexity of the decision logic and the cognitive capabilities of individuals dealing with such complexity.

Furthermore, while the framework demonstrated strengths in providing an understandable and generic process framework, our study also revealed a weakness in terms of actionable recommendations to improve the decision-making process. Despite the framework's ability to formalize and incorporate domain knowledge, its effectiveness in guiding decision-making is limited when the necessary decision logic cannot be derived from the provided data.

In conclusion, this study is the first to integrate domain knowledge through the decision requirements level and employ ML to extract the decision logic level within a process framework for guiding data-driven decision-making projects (see Table 11). By providing a structured framework, this research contributes to the standardization of incorporating domain knowledge in

data-driven decision-making. However, this study does not come without its limitations, which we will discuss in the next section.

8.2 Limitations

First, the generalizability of our results is limited, as we conducted a DS study using small focus groups and focused solely on a single real-world setting. Therefore, the extent to which our findings can be applied to other contexts or organizations may be restricted. However, it is important to note that DMN and CRISP-DM, which form the foundation of DMN-D3M, are generic frameworks and notations. This suggests that DMN-D3M has the potential to be domain-independent, as was also confirmed during our focus groups.

Second, it is important to acknowledge the potential for bias in the evaluations conducted in our study. Participants may have been inclined to provide socially preferable (more positive) answers, which could impact the reliability of our findings. Furthermore, the participation of only six experts in the focus groups is relatively low, which may have implications for the diversity and representativeness of the insights. Additionally, the relative lack of experience of the moderator might have influenced the results to some extent as well.

Third, a limitation of this study is that the data analysts participating in the focus groups had a more limited role, primarily monitoring the project through presentations, while domain experts were actively involved and possessed experience in the specific context. This arrangement may have led to limited familiarity of the data analysts with the details of the project compared to the domain experts, potentially impacting their ability to evaluate the framework and the results during the ex-post evaluation.

Fourth, a limitation of this study is the strong dependence on the data analyst(s) involved in the project, as they have a significant impact on the quality of the implementation. The ability of the analyst to create an understandable DRD diagram was emphasized by domain experts. However, due to the lack of a clearly defined approach for this task, the outcomes may vary across different analysts. In our implementation, we enhanced the readability of the DRD by incorporating boxes around the three high-level input categories. This visual aid helped simplify the graph and facilitated comprehension for domain experts. Additionally, it was observed that the explanations provided during the evaluation phase had a significant influence on the domain experts' understanding of the insights generated from the analysis. This raises questions about whether the framework provides sufficient guidance to assist data analysts in achieving the desired results.

Lastly, it should be noted that the scope of this study specifically focused on ML in a decision-making context, which resulted in a narrow definition of domain knowledge. The domain knowledge considered in this study was limited to the specific decision-making knowledge. Consequently, the findings may not fully capture the broader range of ML applications or domains where different types of domain knowledge may be relevant.

8.3 Future work

We identified several research directions for future studies, both theoretically and practically at RHDHV.

8.3.1 Practical Research Directions at RHDHV

For the specific rail maintenance context at RHDHV, there is a need to explore more accurate and reliable element-level data collection methods. One potential approach is to leverage automated measurement techniques integrated within a measurement train, which could enable more precise and faster data collection. However, it is important to note that this would result in significantly larger volumes of data, necessitating a different skill set within RHDHV to effectively handle and analyze such data. Additionally, implementing this approach would require a comprehensive reorganization of the inspection and decision-making process, which can be a costly undertaking considering the expenses associated with the measurement train and resources required for data analysis. Nonetheless, the integration of such data collection methods leads to more objective data and has the potential to be complemented by an algorithm that determines the optimal budget allocation, thereby enhancing the overall efficiency of the rail maintenance operations. By adopting such data collection methods, the decision-making process can transition towards a more data-driven process that is less dependent of domain experts.

8.3.2 Theoretical Research Directions

First, it would be valuable to investigate the generalizability of our framework beyond the specific context of RHDHV. Understanding its applicability, strengths, weaknesses, and limitations in different real-world settings can provide valuable insights and allow for its further development.

Second, the framework can be further investigated by exploring alternatives to the decision table, particularly in scenarios that involve more complex decision-making. One such alternative is the utilization of black box models integrated with eXplainable Artificial Intelligence (XAI) techniques, which can provide interpretability and transparency to complex models. In our study, domain experts clearly expressed a preference for visualizations, such as feature importances, over the decision logic itself, which can be challenging to interpret. This suggests that XAI explanations may offer greater value than the decision logic alone. However, striking the right balance between the complexity required for accurate decision logic and the need for transparency and understandability is crucial and context-dependent.

Third, an important aspect that has not been fully addressed in our study is the potential conflict between domain knowledge and insights generated from data. Future studies should investigate how to effectively navigate situations where domain knowledge and data-driven insights appear to be incompatible, as well as how to manage conflicting domain knowledge from multiple experts.

A. DMN-D3M generic tasks

Table 20: DMN-D3M **tasks** and their *output*

Decision Requirements Formalization	**Data Understanding**	**Data Preparation**	**Decision Logic Extraction**	**Evaluation**
Elicit domain knowledge *Context* *Requirements* **Formalize decision requirements** *Decision Requirements Diagram (DRD)* **Validate decision requirements** *DRD feedback*	**Determine data storage location(s)** *Data storage location(s)* **Collect raw data** *Raw data* **Describe data** *Data description (e.g. descriptive statistics, data visualizations)*	**Clean data** *Cleaned data* **Engineer features** *Data with correct features* **Integrate data** *Merged data* **Split data** *Splitted data (e.g. train and test split)*	**Select algorithm(s)** *Algorithm(s)* **Extraction decision logic** *Decision rules* *Decision table* *Parameter settings* **Assess performance criteria** *Decision logic performance*	**Evaluate results** *Decision logic performance feedback* *Decision logic meaningfulness feedback*

B. Information Systems Design Theory (ISDT)

Table 21: Framework ISDT

No.	Component	Description
1	Purpose and Scope	The framework is designed to provide guidance for machine learning efforts within the context of data-driven decision-making, focusing on the extraction of decision logic that is implicitly captured in historical data.
2	Constructs	The framework includes the hierarchical process framework itself, comprising three levels of abstraction: phases, generic tasks, and process instances. Additionally, the key roles involved are defined as data analysts and domain experts.
3	Principle of form and function	The framework is designed to prioritize the involvement of domain knowledge throughout the ML process. It integrates Decision Model and Notation (DMN) and the Cross Industry Standard Process for Data Mining (CRISP-DM) as complementary frameworks.
4	Artifact mutability	The framework provides an end-to-end approach that can be customized to suit specific project needs by modifying or selecting particular phases. For example, an anticipated adaptation is the use of black-box models in scenarios where the decision logic exceeds the capabilities of decision tables.
5	Testable propositions	1. The use of DMN's Decision Requirements Diagram (DRD) as a formalization technique improves communication and understanding among stakeholders in the Decision Requirements Formalization phase. 2. The collaboration between data analysts and domain experts in the Data Understanding phase improves the understanding of the data and enhances the quality of insights generated. 3. The use of ML-based decision tables in the Decision Logic Extraction phase results in decision logic that is interpretable and maintainable.
6	Justifactory knowledge	The framework draws upon existing frameworks such as DMN and CRISP-DM as its theoretical foundations. These frameworks provide the rationale for the design decisions and offer established approaches in ML and decision modeling.
7	Principles of implementation	The framework includes a high-level description of phases and corresponding generic activities for its implementation, while allowing for flexibility at the process instance level.
8	Expository instantiation	The framework is illustrated in a real-world context.

C. Individual component grades

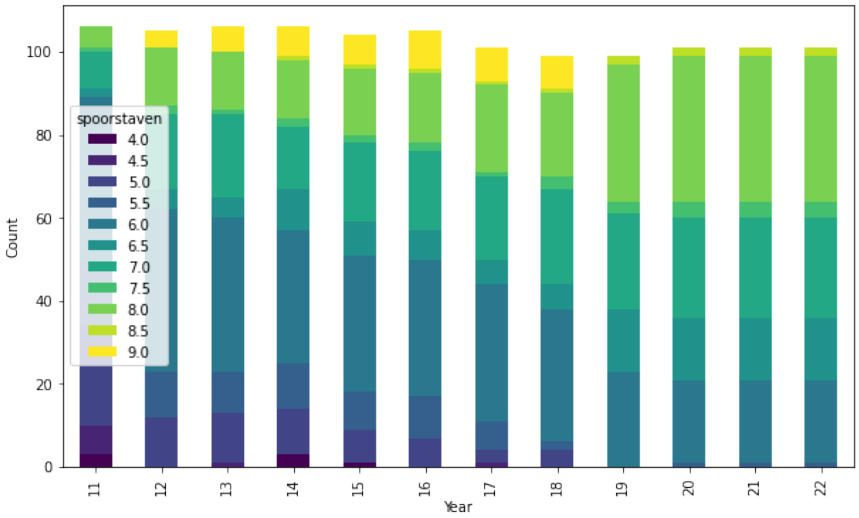


FIGURE 19: Rail Grades by Year

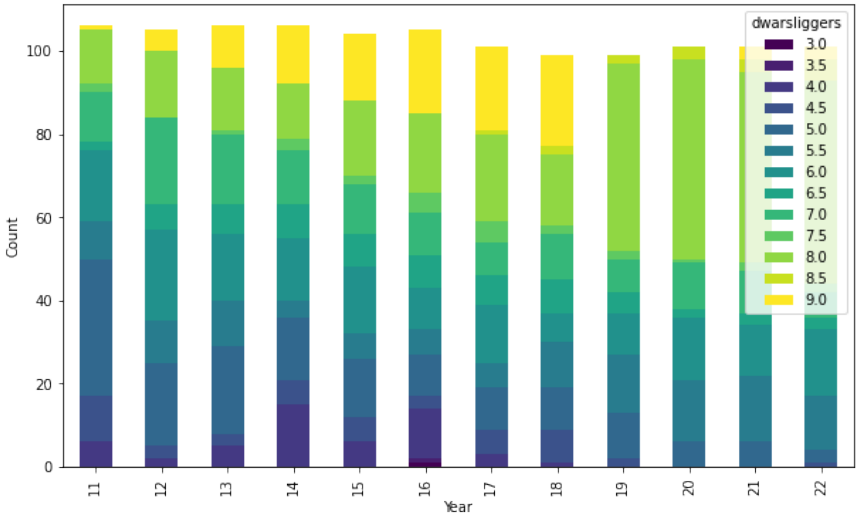


FIGURE 20: Sleeper Grades by Year

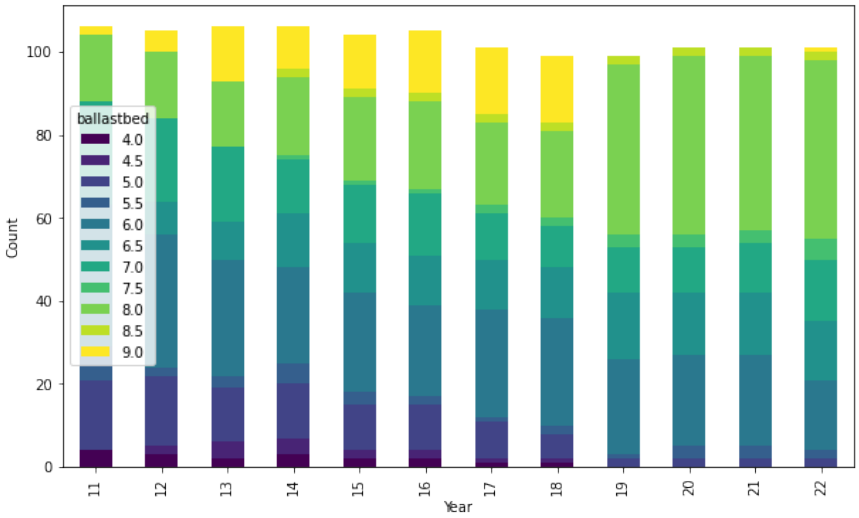


FIGURE 21: Ballast Grades by Year

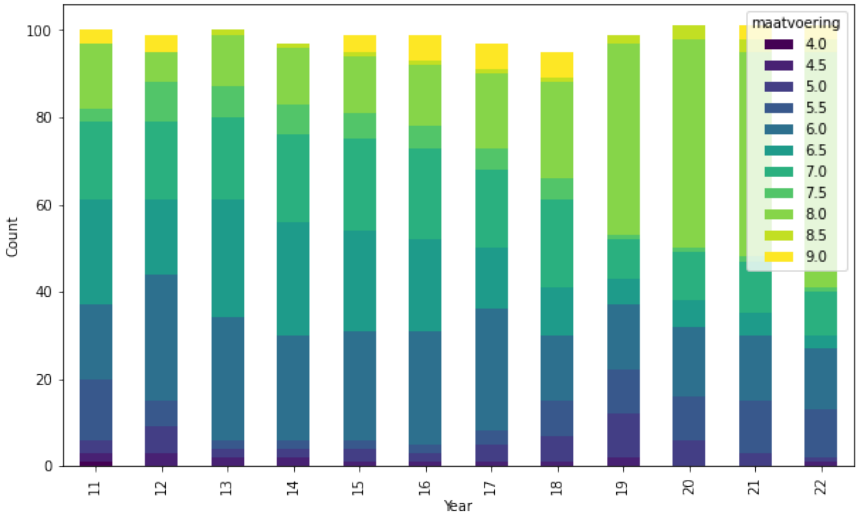


FIGURE 22: Dimensions Grades by Year

D. Results other scenarios

D.1 Scenario 2

To predict whether maintenance should be performed the next year or not, we followed the same procedure as depicted in scenario 1. The results of the CART hyperparameter optimization are shown in Figure 23.

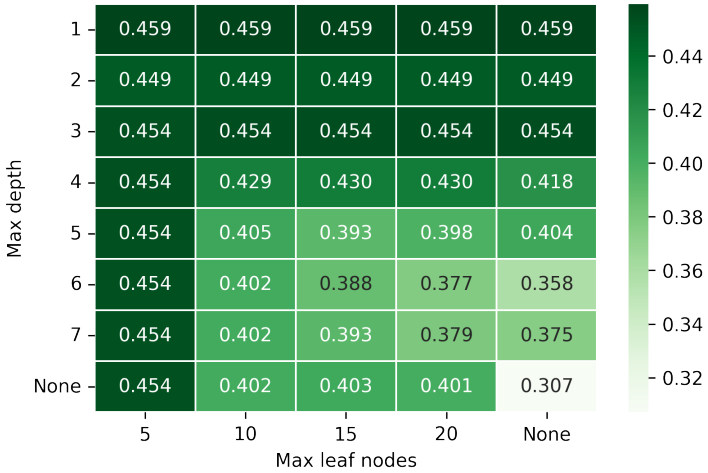


FIGURE 23: CART hyperparameter optimization (Scenario 2)

CART obtains a maximum train F1-score of only 0.459 and a corresponding test set F1-score of 0.487. For C4.5, we observed heavy overfitting with a train score of 0.656 and a test score of only 0.193. These results indicate that neither model performs well enough to be considered useful in practical scenarios.

D.2 Scenario 3

In scenario 3, we used micro-averaging for the F1-score in this multi-class scenario, which entails simply aggregating all predictions, rather than evaluating each class separately. Unfortunately, both CART and C4.5 models exhibit poor performance, making them unsuitable for real-world applications. With a maximum train F1-score of 0.266 and corresponding test set F1-score of 0.304, the results of CART are underwhelming (Figure 24). C4.5 again performs better on the training set (0.416), but does not generalize to the test set (0.275).

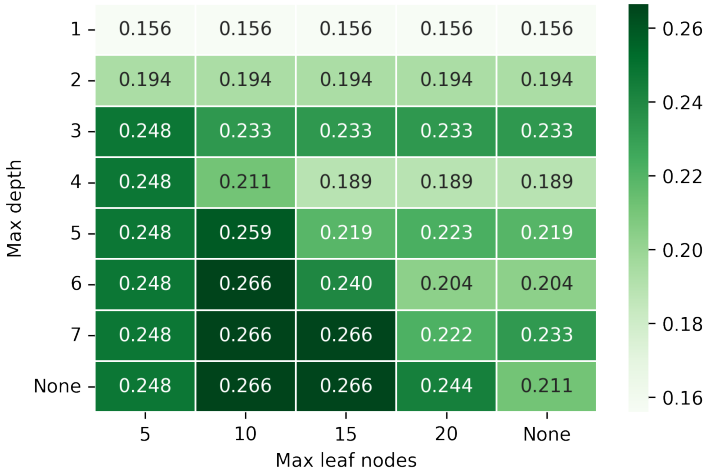


FIGURE 24: CART hyperparameter optimization (Scenario 3)

E. Extended results

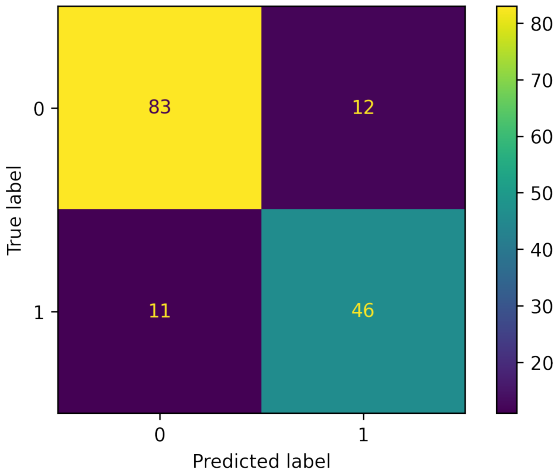


FIGURE 25: Confusion matrix

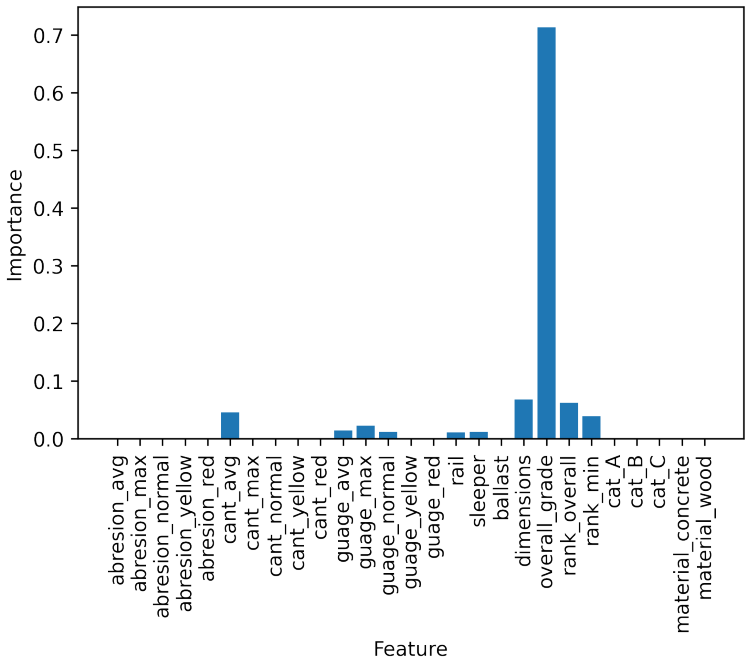


FIGURE 26: Feature importance

Table 22: Decision table

U	Overall grade	Dimensions	Rank (overall grade)	Cant (avg)	Rank (min grade)	Gauge (max)	Gauge (avg)	Sleeper	Gauge (normal)	Rail	Decision
1	≤6.42	-	-	-	≤16.50	-	≤33.93	-	-	-	Not in MIP
2	≤6.42	-	-	-	[16.50-63.50]	-	≤33.93	-	-	-	In MIP
3	≤6.42	-	-	-	≤63.50	-	>33.93	-	-	-	In MIP
4	≤6.42	-	-	-	>63.50	-	-	-	-	-	Not in MIP
5	[6.42-6.81]	-	-	≤4.09	-	-	-	-	-	≤5.50	In MIP
6	[6.42-6.81]	-	-	≤4.09	-	-	-	-	-	>5.50	Not in MIP
7	[6.42-6.81]	-	≤32.00	>4.09	-	-	-	-	-	-	Not in MIP
8	[6.42-6.81]	-	>32.00	≤10.92	-	-	-	-	-	-	In MIP
9	[6.42-6.81]	-	>32.00	>10.92	-	-	-	-	-	-	Not in MIP
10	[6.81-6.94]	≤6.75	-		-	-	-	-	-	-	Not in MIP
11	>6.81	[6.75-7.75]	≤61.50	-	-	≤45.50	-	-	-	-	In MIP
12	>6.81	[6.75-7.75]	>61.50	-	-	≤45.50	-	-	-	-	Not in MIP
13	>6.81	[6.75-7.75]	-	-	-	>45.50	-	-	-	-	Not in MIP
14	>6.81	>7.75	-	-	-	-	-	≤6.75	-	-	In MIP
15	>6.81	>7.75	-	-	-	-	-	>6.75	-	-	Not in MIP
16	>6.94	≤6.75	-	-	-	-	-	-	≤7.50	-	Not in MIP
17	>6.94	≤6.75	-	-	≤69.00	-	-	-	>7.50	-	In MIP
18	>6.94	≤6.75	-	-	>69.00	-	-	-	>7.50	-	Not in MIP

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, *6*, 52138–52160.

Aier, S., & Fischer, C. (2011). Criteria of progress for information systems design theories. *Information Systems and E-Business Management*, *9*, 133–172.

Akter, S., Bandara, R., Hani, U., Wamba, S. F., Foropon, C., & Papadopoulos, T. (2019). Analytics-based decision-making for service systems: A qualitative study and agenda for future research. *International Journal of Information Management*, *48*, 85–95.

Arco, L., Nápoles, G., Vanhoenshoven, F., Lara, A. L., Casas, G., & Vanhoof, K. (2021). Natural language techniques supporting decision modelers. *Data Mining and Knowledge Discovery*, *35*(1), 290–320.

Baesens, B., Setiono, R., Mues, C., & Vanthienen, J. (2003). Using neural network rule extraction and decision tables for credit-risk evaluation. *Management science*, *49*(3), 312–329.

Bazhenova, E., Buelow, S., & Weske, M. (2016). Discovering decision models from event logs. In *International conference on business information systems* (pp. 237–251).

Bazhenova, E., & Weske, M. (2016). Deriving decision models from process models by enhanced decision mining. In *International conference on business process management* (pp. 444–457).

Bork, D., Ali, S. J., & Dinev, G. M. (2023). Ai-enhanced hybrid decision management. *Business & Information Systems Engineering*, *65*(2), 179–199.

Burnham, J. F. (2006). Scopus database: a review. *Biomedical digital libraries*, *3*(1), 1–8.

Car, N. J. (2018). Using decision models to enable better irrigation decision support systems. *Computers and Electronics in Agriculture*, *152*, 290–301.

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, *8*(8), 832.

Chiheb, F., Boumahdi, F., & Bouarfa, H. (2019a). A conceptual model for describing the integration of decision aspect into big data. *International Journal of Information System Modeling and Design (IJISMD)*, *10*(4), 1–23.

Chiheb, F., Boumahdi, F., & Bouarfa, H. (2019b). A new model for integrating big data into phases of decision-making process. *Procedia Computer Science*, *151*, 636–642.

Craven, M., & Shavlik, J. (1995). Extracting tree-structured representations of trained networks. *Advances in neural information processing systems*, *8*.

Deng, C., Ji, X., Rainey, C., Zhang, J., & Lu, W. (2020). Integrating machine learning with human knowledge. *Iscience*, *23*(11), 101656.

De Smedt, J., Hasić, F., vanden Broucke, S. K., & Vanthienen, J. (2019). Holistic discovery of decision models from process execution data. *Knowledge-Based Systems*, *183*, 104866.

Etikala, V., & Vanthienen, J. (2021). An overview of methods for acquiring and generating decision models. In *International conference on knowledge science, engineering and management* (pp. 200–208).

Etikala, V., Van Veldhoven, Z., & Vanthienen, J. (2020). Text2dec: extracting decision dependencies from natural language text for automated dmn decision modelling. In *International conference on business process management* (pp. 367–379).

Etinger, D., Simić, S. D., & Buljubašić, L. (2019). Automated decision-making with dmn: from decision trees to decision tables. In *2019 42nd international convention on information and communication technology, electronics and microelectronics (mipro)* (pp. 1309–1313).

Goossens, A., De Smedt, J., & Vanthienen, J. (2023). Extracting decision model and notation models from text using deep learning techniques. *Expert Systems with Applications*, *211*, 118667.

Gregor, S. (2009). Building theory in the sciences of the artificial. In *Proceedings of the 4th international conference on design science research in information systems and technology* (pp. 1–10).

Gregor, S., & Jones, D. (2007). The anatomy of a design theory. *Journal of the Association for Information Systems*, *8*(5), 1.

Guz, A. N., & Rushchitsky, J. (2009). Scopus: A system for the evaluation of scientific journals. *International Applied Mechanics*, *45*(4), 351–362.

Hasić, F., & Vanthienen, J. (2019). Complexity metrics for dmn decision models. *Computer Standards & Interfaces*, *65*, 15–37.

Hevner, Chatterjee, S., Tremblay, M. C., Hevner, A. R., & Berndt, D. J. (2010). The use of focus groups in design science research. *Design Research in Information Systems: Theory and Practice*, 121–143.

Hevner, March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *Management Information Systems Quarterly*, *28*(1), 6.

Huysmans, J., Dejaeger, K., Mues, C., Vanthienen, J., & Baesens, B. (2011). An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems*, *51*(1), 141–154.

Kerrigan, D., Hullman, J., & Bertini, E. (2021). A survey of domain knowledge elicitation in applied machine learning. *Multimodal Technologies and Interaction*, *5*(12), 73.

Kopanas, I., Avouris, N. M., & Daskalaki, S. (2002). The role of domain knowledge in a large scale data mining project. In *Methods and applications of artificial intelligence: Second hellenic conference on ai, setn 2002 thessaloniki, greece, april 11–12, 2002 proceedings 2* (pp. 288–299).

Li, Y., Zhang, H., Roy, U., & Lee, Y. T. (2017). A data-driven approach for improving sustainability assessment in advanced manufacturing. In *2017 ieee international conference on big data (big data)* (pp. 1736–1745).

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*.

Lima, E., Mues, C., & Baesens, B. (2009). Domain knowledge integration in data mining using decision tables: case studies in churn prediction. *Journal of the Operational Research Society*, *60*(8), 1096–1106.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision support systems*, *15*(4), 251–266.

Martens, D., Baesens, B., Van Gestel, T., & Vanthienen, J. (2007). Comprehensible credit scoring models using rule extraction from support vector machines. *European journal of operational research*, *183*(3), 1466–1476.

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of web of science and scopus: a comparative analysis. *Scientometrics*, *106*(1), 213–228.

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., & Müller, K.-R. (2017). Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern recognition*, *65*, 211–222.

Mues, C., Baesens, B., Setiono, R., & Vanthienen, J. (2005). From knowledge discovery to implementation: A business intelligence approach using neural network rule extraction and decision tables. In *Biennial conference on professional knowledge management/wissensmanagement* (pp. 483–495).

Mylopoulos, J. (1992). Conceptual modelling and telos. *Conceptual modelling, databases, and CASE: An integrated view of information system development*, 49–68.

OMG. (2015). Decision model and notation, version 1.0. `https://www.omg.org/spec/DMN/1.0`.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of management information systems*, *24*(3), 45–77.

Prat, N., Comyn-Wattiau, I., & Akoka, J. (2014). Artifact evaluation in information systems design-science research–a holistic view.

Pries-Heje, J., Baskerville, R., & Venable, J. (2008). Strategies for design science research evaluation. In *Conference proceedings, 16th european conference on information systems.*

Quinlan, J. R. (1993). *C4. 5: programs for machine learning.* Morgan Kaufmann, Chambery, France.

Quishpi, L., Carmona, J., & Padró, L. (2021). Extracting decision models from textual descriptions of processes. In *International conference on business process management* (pp. 85–102).

Saltz, J., Hotz, N., Wild, D., & Stirling, K. (2018). Exploring project management methodologies used within data science teams.

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying crisp-dm process model. *Procedia Computer Science*, *181*, 526–534.

Servadei, L., Schmidt, R., & Bär, F. (2016). Artificial neural network for supporting medical decision making: a decision model and notation approach to spondylolisthesis and disk hernia. In *Otm confederated international conferences" on the move to meaningful internet systems"* (pp. 217–227).

Setiono, R., & Liu, H. (1996). Symbolic representation of neural networks. *Computer*, *29*(3), 71–77.

Setiono, R., & Liu, H. (1997). Neurolinear: From neural networks to oblique decision rules. *Neurocomputing*, *17*(1), 1–24.

Sharma, S., & Osei-Bryson, K.-M. (2008). Organization-ontology based framework for implementing the business understanding phase of data mining projects. In *Proceedings of the 41st annual hawaii international conference on system sciences (hicss 2008)* (pp. 77–77).

Simić, S., Tanković, N., & Etinger, D. (2020). Automated decision modeling with dmn and bpmn: A model ensemble approach. *Advances in Intelligent Systems and Computing*, *1026*, 789-794.

Sonnenberg, C., & Vom Brocke, J. (2012). Evaluations in the science of the artificial–reconsidering the build-evaluate pattern in design science research. In *Design science research in information systems. advances in theory and practice: 7th international conference, desrist 2012, las vegas, nv, usa, may 14-15, 2012. proceedings 7* (pp. 381–397).

Stefanou, C. J. (2001). A framework for the ex-ante evaluation of erp software. *European Journal of Information Systems*, *10*(4), 204–215.

Steinberg, D. (2009). Cart: classification and regression trees. In *The top ten algorithms in data mining* (pp. 193–216). Chapman and Hall/CRC.

Stewart, D. W., Shamdasani, P. N., & Rook, D. W. (2007). *Focus groups: Theory and practice* (2nd ed., Vol. 20). Newbury Park, CA: Sage Publications.

Varshney, K. R., Khanduri, P., Sharma, P., Zhang, S., & Varshney, P. K. (2018). Why interpretability in machine learning? an answer using distributed detection and data fusion theory. *arXiv preprint arXiv:1806.09710*.

Venable, J., Pries-Heje, J., & Baskerville, R. (2012). A comprehensive framework for evaluation in design science research. In *Design science research in information systems. advances in theory and practice: 7th international conference, desrist 2012, las vegas, nv, usa, may 14-15, 2012. proceedings 7* (pp. 423–438).

Venable, J., Pries-Heje, J., & Baskerville, R. (2016). Feds: a framework for evaluation in design science research. *European journal of information systems*, *25*, 77–89.

Von Rueden, L., Mayer, S., Beckh, K., Georgiev, B., Giesselbach, S., Heese, R., ... others (2021). Informed machine learning–a taxonomy and survey of integrating prior knowledge into learning systems. *IEEE Transactions on Knowledge and Data Engineering*, *35*(1), 614–633.

Wagner, W. P. (2017). Trends in expert system development: A longitudinal content analysis of over thirty years of expert system case studies. *Expert systems with applications*, *76*, 85–96.

Wets, G., Vanthienen, J., & Timmermans, H. (1998). Modelling decision tables from data. In *Research and development in knowledge discovery and data mining, second pacific-asia conference, pakdd-98, melbourne, australia, april 15–17, 1998, proceedings* (pp. 412–13).

Wirth, R., & Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (Vol. 1, pp. 29–39).